Robust Minimal Instability of the Top Trading Cycles Mechanism[*]

Battal Doğan[†] Lars Ehlers[‡]

December 13, 2020

Abstract

In the context of priority-based allocation of objects, we formulate methods to compare assignments in terms of their stability. We introduce three basic properties that a reasonable stability comparison should satisfy. We show that, for any stability comparison satisfying the three properties, the top trading cycles mechanism is minimally unstable among *efficient* and *strategy-proof* mechanisms when objects have unit capacities. Our unifying approach covers basically all natural stability comparisons and establishes the robustness of a recent result by Abdulkadiroğlu et al. (2020). When objects have non-unit capacities, we characterize the capacity-priority structures for which our result is preserved.

1 Introduction

Many resource-allocation problems include objects, such as houses, offices, jobs, or school seats, endowed with priority orderings over agents, and a mechanism elicits agents' preferences and allocates the objects based on the preferences and the priorities. In such problems,

[*] Battal Doğan acknowledges financial support from the British Academy/Leverhulme Trust (SRG1819\190133). Lars Ehlers acknowledges financial support from the SSHRC (Canada) and the FRQSC (Québec). We are grateful to four anonymous referees, the Editor, and participants at several seminars and conferences for valuable comments and helpful suggestions.

[†] Department of Economics, University of Bristol, 3B13 The Priory Road Complex, Priory Road, Clifton, BS8 1TU. Email: battal.dogan@bristol.ac.uk.

[‡] Département de Sciences Économiques and CIREQ, Université de Montréal, Montréal, QC H3C 3J7, Canada. Email: lars.ehlers@umontreal.ca.

respecting preferences is captured by the efficiency property, which requires that there is no other assignment at which an agent is better off while no agent is worse off. On the other hand, respecting priorities is captured by the stability property, which requires that there is no "blocking pair" of an agent and an object such that the agent prefers the object to his assignment and he has a higher priority than another agent who receives that object. Unfortunately, no mechanism can guarantee *efficiency* and *stability* at the same time: there exist problems without an assignment that is both *efficient* and *stable*.[1]

In their seminal paper, Abdulkadiroğlu and Sönmez (2003) propose to use the deferred-acceptance (DA) mechanism or the top trading cycles (TTC) mechanism in the context of school choice, depending on whether you want to guarantee *stability* or *efficiency*. Both mechanisms are *strategy-proof*: for each agent, it is a weakly dominant strategy to report his preferences truthfully. On the other hand, the DA mechanism is stable but inefficient whereas the TTC mechanism is *efficient* but unstable. However, the DA mechanism is "constrained *efficient*" as it chooses the agents-optimal (with respect to the Pareto dominance comparison) stable assignment (Gale and Shapley, 1962). Two natural questions arise: (1) What are methods to compare assignments—possibly two unstable assignments—in terms of their stability? (2) Is the TTC mechanism "minimally unstable" with respect to any natural stability comparison?

We address these questions by introducing three basic properties that any reasonable stability comparison should satisfy. The first property *stability-preferred* imposes the following requirement: any stable assignment should be strictly more stable than any unstable assignment. The second property *separability* imposes a requirement only on the domain of problems where each object has unit capacity and formalizes the following idea. Suppose that an assignment μ is more stable than another assignment ν, while for a set of agents and their assigned objects, ν is stable but μ is unstable. Then, μ should be strictly more stable than ν when restricted to the other agents and objects (to be able to justify that μ is *overall* more stable than ν). The third property *consistency* imposes a requirement also only on the domain of problems where each object has unit capacity and formalizes the following idea. Suppose that an assignment μ is more stable than another assignment ν while for a set of agents, μ and ν assignments coincide and these agents and their assigned objects are included in blocking pairs only among themselves at both μ and ν. Then, μ should still be more stable than ν when restricted to the other agents and objects (since the removal of a part where the two assignments coincide should not affect the overall stability comparison).

[1] This follows from an example in Roth (1982). It is more explicitly shown in Abdulkadiroğlu and Sönmez (2003).

Our main result, Theorem 1, states that given any stability comparison satisfying *stability-preferred*, *separability*, and *consistency*, the TTC mechanism is minimally unstable among *efficient* and *strategy-proof* mechanisms when all objects have unit capacities, i.e., there is no other *efficient* and *strategy-proof* mechanism that is more stable than the TTC mechanism with respect to any stability comparison satisfying the three properties.

Our paper is not the first one to address these questions. In a recent paper, Abdulkadiroğlu et al. (2020) propose to compare assignments in terms of their stability by comparing the sets of blocking pairs at these assignments, and calling an assignment *more stable* than another assignment if the set of blocking pairs in the former assignment is a subset of the set of blocking pairs in the latter assignment. Using this natural stability comparison, they show that the TTC mechanism is minimally unstable among *efficient* and *strategy-proof* mechanisms when each object has unit capacity, establishing the first "minimal instability" result for the TTC mechanism in the literature.[2] The stability comparison in Abdulkadiroğlu et al. (2020) satisfies our three properties and therefore their result follows as a corollary to ours. Moreover, our result shows that the TTC mechanism is minimally unstable when each object has unit capacity with respect to many other—from our axiomatic perspective, to all—natural stability comparisons. For example, a natural alternative is to count the number of blocking pairs, which induces a complete comparison method (as all assignments can be compared by counting blocking pairs). One may also consider comparison methods that are not based on the set of blocking pairs, but based on alternative sets such as the set of blocking triplets as in Kwon and Shorrer (2019),[3] or the set of blocking agents as in **?**.[4]

On the technical front, our main proof arguments are considerably different than the corresponding ones of Abdulkadiroğlu et al. (2020).[5] A key step in the proof is to show that any *efficient* and *strategy-proof* mechanism that is more stable than the TTC mechanism must satisfy a *mutual-best property*: if an agent and an object mutually top-rank each other

[2] Abdulkadiroğlu et al. (2020) use the "justified envy minimality" terminology instead of "minimal instability". In the context of our paper, which is the same as the context of Abdulkadiroğlu et al. (2020), stability has a fairness interpretation and a blocking pair is equivalent to an instance of justified envy. In a recent paper, Romm et al. (2020) show that in different contexts, the concepts of blocking and justified envy may diverge.

[3] A blocking triplet includes, in addition to a blocking pair, an agent who violates the priority of the agent in the blocking pair. Kwon and Shorrer (2019) show that TTC mechanism is minimally unstable among *efficient* and *strategy-proof* mechanisms in one-to-one matching when stability comparison is based on comparing (in the set-inclusion sense) sets of blocking triplets.

[4] A blocking agent is an agent who is involved in at least one blocking pair. In **?**, we drop *strategy-proofness* and investigate *efficient* and *minimally unstable* Pareto improvements over the deferred acceptance mechanism for several natural stability comparisons.

[5] They are also different from the main arguments in the characterizations of the TTC mechanism in different contexts, such as by Ma (1994), Svensson (1999), **?**, and **?**.

in their preference and priority orderings, then the agent should receive the object.[6] If the stability comparison is based on comparing the sets of blocking pairs in the set inclusion sense as in Abdulkadiroğlu et al. (2020), then the mutual-best property is immediate: suppose that the agent is not assigned the object; then, they constitute a blocking pair, while they are not a blocking pair under the TTC mechanism, contradicting that the mechanism is more stable than the TTC mechanism. Such a conclusion is not immediate if the stability comparison is, for example, based on counting blocking pairs. In Section 4 after stating Theorem 1, we provide a detailed sketch of the proof.

In some applications such as school choice, objects do not have unit capacities, and Theorem 1 fails. In fact, it follows from Example 1 in Abdulkadiroğlu et al. (2020) that for any stability comparison that satisfies *stability-preferred*, the TTC mechanism is not minimally unstable among *efficient* and *strategy-proof* mechanisms.[7] In Theorem 2, when each object has non-unit capacity we characterize the capacity-priority structures for which the TTC mechanism is robustly minimally unstable among *efficient* and *strategy-proof* mechanisms. Our result reveals that the TTC mechanism is not robustly minimally unstable when each object has at least two copies,[8] except for the capacity-priority structures for which the TTC mechanism is always stable, in which case it is trivially robustly minimally unstable.

Our paper is related to the literature on understanding the implications of *efficiency* and *strategy-proofness* in object allocation such as, among others, Pápai (2000), Pycia and Ünver (2017), Kesten (2010), and Kesten and Kurino (2019). Another related paper is Bonkoungou and Nesterov (2020) who use natural stability comparisons to explain some school choice reforms. Finally, although the investigation of stability comparisons in two-sided matching is new, there is a related literature on stability comparisons for roommates problems such as, among others, Abraham et al. (2005).[9]

The paper is organized as follows. Section 2 introduces priority-based object allocation problems. Section 3 defines stability comparison methods, introduces basic properties for stability comparisons, and provides examples of natural stability comparisons. Section 4

[6] To our knowledge, *mutual-best property* was first studied in **?** in this context.

[7] For the multi-capacity case, Abdulkadiroğlu et al. (2020) provide a justification for the TTC mechanism from a different perspective and show that the TTC mechanism outperforms *serial dictatorship*, an obvious *efficient* alternative, by admitting fewer blocking pairs in an average sense when every possible priority profile is considered or when participants' priorities are drawn uniform randomly. Note that this justification has a cardinal nature, and we believe that incorporating stability comparisons with cardinal nature complements this alternative justification.

[8] In fact, not minimally unstable for any stability comparison that satisfies *stability-preferrred*.

[9] Abraham et al. (2005) define *almost stable matchings* as matchings that minimize the number of blocking pairs, which is the *roommates problem counterpart* of the blocking pairs cardinality comparison considered in this paper and also in **?**.

defines the TTC mechanism and states our main results. Section 5 concludes. The Appendix contains the proofs relegated from the main text.

2 The Model

Let $\mathcal{N}$ denote the set of potential agents and $\mathcal{C}$ denote the set of potential objects. We call a tuple $(N, C, R, q, \succeq)$ a **problem**, where

- $N \subset \mathcal{N}$ is a finite set of agents,
- $C \subset \mathcal{C}$ is a finite set of objects,
- $R = (R_i)_{i \in N}$ is a preference profile, which is a profile of linear orderings over $C \cup \{\emptyset\}$ where $\emptyset$ represents an outside option for the agent,[10]
- $q = (q_c)_{c \in C}$ is a capacity profile with $q_c \in \mathbb{N}$ for each $c \in C$,
- $\succeq = (\succeq_c)_{c \in C}$ is a priority profile, which is a profile of linear orderings over N.[11]

The strict part of a preference ordering R_i is denoted by P_i.[12] Object c is **acceptable** to agent i if he prefers it to his outside option, i.e., $c \; P_i \; \emptyset$. Object c has **unit capacity** if $q_c = 1$, and otherwise object c has **non-unit capacity**. The strict part of a priority ordering $\succeq_c$ is denoted by $\succ_c$. Let $\mathcal{E}^{(N,C)}$ denote the set of all problems (or economies) including N as the set of agents and C as the set of objects, and let $\mathcal{E}$ denote the set of all problems including any finite sets of agents and objects.

Given a problem $E = (N, C, R, q, \succeq) \in \mathcal{E}$, a set of agents $N' \subseteq N$ and objects $C' \subseteq C$, we call $E|_{(N',C')}$ as the **restriction** of E to (N', C'), where $E|_{(N',C')}$ is obtained from E by simply removing $N \backslash N'$ and $C \backslash C'$, and also removing them from q, R, and $\succeq$ while keeping relative orderings of the remaining agents and the relative orderings and capacities of the remaining objects the same.

Given a problem $E = (N, C, R, q, \succeq) \in \mathcal{E}$, an assignment is a mapping $\mu : N \cup C \to N \cup C \cup \{\emptyset\}$ such that

[10] Formally, a preference ordering is a complete, transitive, and anti-symmetric binary relation over $C \cup \{\emptyset\}$. Binary relation R_i over $C \cup \{\emptyset\}$ is *complete* if, for every $c_1, c_2 \in C \cup \{\emptyset\}$, $c_1 R_i c_2$ or $c_2 R_i c_1$. It is *transitive* if, for every $c_1, c_2, c_3 \in C \cup \{\emptyset\}$, $c_1 R_i c_2$ and $c_2 R_i c_3$ imply $c_1 R_i c_3$. It is *anti-symmetric* if, for every $c_1, c_2 \in C \cup \{\emptyset\}$, $c_1 R_i c_2$ and $c_2 R_i c_1$ imply $c_1 = c_2$.

[11] Formally, a priority ordering is a complete, transitive, and anti-symmetric binary relation over N.

[12] That is, if $c_1, c_2 \in C \cup \{\emptyset\}$, $c_1 \neq c_2$, and $c_1 \; R_i \; c_2$, then $c_1 \; P_i \; c_2$.

(i) for each $i \in N$, $\mu(i) \in C \cup \{\emptyset\}$,

(ii) for each $c \in C$, $\mu(c) \subseteq N$ such that $|\mu(c)| \leq q_c$, and

(iii) for each $i \in N$ and each $c \in C$, $i \in \mu(c)$ if and only if $c = \mu(i)$.

Let $\mathcal{A}(E)$ denote the set of all possible assignments at the problem E. Note that $\mathcal{A}(E)$ is determined by (N, C, q).

An assignment μ **Pareto dominates** another assignment μ' if for each $i \in N$, $\mu(i)\ R_i\ \mu'(i)$ and there exists $i \in N$ such that $\mu(i)\ P_i\ \mu'(i)$. An assignment μ is ***efficient*** if it is not Pareto dominated. Note that *efficiency* implies ***individual rationality***: for each $i \in N$, $\mu(i)\ R_i\ \emptyset$.

A pair $(i, c) \in N \times C$ **blocks** μ if $c\ P_i\ \mu(i)$ and $[|\mu(c)| < q_c$ or there exists $j \in \mu(c)$ such that $i \succ_c j]$. Let

$$B(\mu) = \{(i, c) \in N \times C : (i, c) \text{ blocks } \mu\}$$

denote the set of blocking pairs at μ. In addition, for each $i \in N$, let $B_i(\mu) = \{c \in C : (i, c) \in B(\mu)\}$ denote the set of objects together with which agent i constitute a blocking pair, and for each $c \in C$, $B_c(\mu) = \{i \in N : (i, c) \in B(\mu)\}$ denote the set of agents together with whom object c constitute a blocking pair.

An assignment μ is **stable** if it is individually rational and includes no blocking pair. Unfortunately, there exist problems without an assignment that is both *efficient* and *stable* (Roth, 1982).

A **mechanism** associates each problem with an assignment. When we say that a mechanism satisfies a certain assignment property, such as *efficiency*, we mean that at each problem, the assignment prescribed by the mechanism satisfies the property.

A mechanism φ is ***strategy-proof*** if reporting true preferences is a weakly dominant strategy for each agent in the preference revelation game induced by φ, that is, for each problem $(N, C, R, q, \succeq)$, each $i \in N$ and each preference ordering R'_i,

$$\varphi_i(N, C, R, q, \succeq)\ R_i\ \varphi_i(N, C, (R'_i, R_{-i}), q, \succeq).$$

When $(N, C, q, \succeq)$ is clear, we often denote a problem simply by its preference profile R. Now, using our convention, the above simply says $\varphi_i(R)\ R_i\ \varphi_i(R'_i, R_{-i})$.

Remark 1 *Given an assignment μ, i has **justified envy** towards j if $\mu(j) = c\ P_i\ \mu(i)$ and*

$i \succ_c j$. Note that any efficient assignment is stable if and only if it contains no justified envy (i.e., no agent has justified envy towards any other agent).

3 A Unifying Approach to Stability Comparisons

A **stability comparison** is a function f associating with each problem $E \in \mathcal{E}$ a binary relation over assignments $\succsim_f^E \subseteq \mathcal{A}(E) \times \mathcal{A}(E)$. We use the convention and write $\mu \succsim_f^E \nu$ instead of $(\mu, \nu) \in \succsim_f^E$, and $\mu \succnsim_f^E \nu$ instead of $[\mu \succsim_f^E \nu$ and not $\nu \succsim_f^E \mu]$. We read $\mu \succsim_f^E \nu$ as "μ is f-more stable than ν at E" and $\mu \succnsim_f^E \nu$ as "μ is strictly f-more stable than ν at E". Note that we do not impose any structure on a stability comparison (such as completeness or transitivity).[13] Later we will describe several examples of stability comparison methods. Also note that, when (N, C, q) is fixed, although the set of assignments is independent of the preference or the priority profile, the stability comparison may vary with the preference and the priority profile, that is, stability comparisons depend on the whole problem.

We now introduce three basic properties that a reasonable stability comparison should satisfy.[14]

3.1 Stability-preferred

The first property imposes the following requirement: any stable assignment should be strictly more stable than any unstable assignment.

Stability-preferred: For each $E \in \mathcal{E}$ and $\mu, \nu \in \mathcal{A}(E)$, if $B(\mu) = \emptyset \neq B(\nu)$, then $\mu \succnsim_f^E \nu$.

3.2 Separability

The second property *separability* imposes a requirement only on the domain of problems where each object has unit capacity. It formalizes the following idea. Suppose that an assignment μ is more stable than another assignment ν, while for a set of agents and their assigned objects, ν is stable but μ is unstable. Then, μ should be strictly more stable than ν when restricted to the other agents and objects (to be able to justify that μ is *overall* more stable than ν).

[13] Here (i) $\succsim_f^E$ is complete if for all $\mu, \nu \in \mathcal{A}(E)$ we have $\mu \succsim_f^E \nu$ or $\nu \succsim_f^E \mu$ and (ii) $\succsim_f^E$ is transitive if $\mu \succsim_f^E \nu$ and $\nu \succsim_f^E \eta$ imply $\mu \succsim_f^E \eta$.

[14] We show in Appendix B that none of the three properties is implied by the other two.

More precisely, suppose that μ is at least as stable as ν and also that there is a set of agents N' whose aggregate assignments[15] are the same at μ and ν, i.e., $\mu(N') = \nu(N') = C'$. Suppose also that no agent in N' is involved in a blocking pair at ν and also no object in C' is involved in a blocking pair at ν, while an agent in N' and an object in C' constitute a blocking pair at μ. Then, when restricted to the other agents and objects $(N \setminus N', C \setminus C')$, μ should be strictly more stable than ν.

We next provide a formal definition of *separability.* Given a unit-capacity problem and a set of agents N' with $\mu(N') = C'$, let $\mu|_{N\setminus N'}$ denote the **restriction** of μ to $N\setminus N'$ and $C\setminus C'$, where $\mu|_{N\setminus N'}$ is obtained from μ by simply removing N' and C' while keeping the assignments of $N\setminus N'$ the same as in μ. Note that $\mu|_{N\setminus N'} \in \mathcal{A}(E|_{(N\setminus N', C\setminus C')})$.

Separability: For each unit-capacity problem $E \in \mathcal{E}$, each pair of assignments $\mu, \nu \in \mathcal{A}(E)$ such that $\mu \gtrsim_f^E \nu$, and each pair of (N', C') such that $\mu(N') = \nu(N') = C'$, if $B_x(\nu) = \emptyset$ for each $x \in N' \cup C'$ and $(i, c) \in B(\mu)$ for some $(i, c) \in N' \times C'$, then $\mu|_{N\setminus N'} \gnsim_f^{E'} \nu|_{N\setminus N'}$, where $E' = E|_{(N\setminus N', C\setminus C')}$.

3.3 Consistency

The third property *consistency* imposes a requirement also only on the domain of problems where each object has unit capacity. It formalizes the following idea. Suppose that an assignment μ is more stable than another assignment ν while for a set of agents, μ and ν assignments coincide and these agents and their assigned objects are included in blocking pairs only among themselves at both μ and ν. Then, μ should still be more stable than ν when restricted to the other agents and objects (since the removal of a part where the two assignments coincide should not affect the overall stability comparison).

We next provide a formal definition of *consistency.*

Consistency: For each unit-capacity problem $E \in \mathcal{E}$, each pair of assignments $\mu, \nu \in \mathcal{A}(E)$ such that $\mu \gtrsim_f^E \nu$, and each $\emptyset \neq N' \subseteq N$ such that $\nu(i) = \mu(i)$ for all $i \in N'$, if $B_i(\mu) = B_i(\nu) \subseteq \mu(N') = \nu(N') = C'$ for all $i \in N'$ and $B_c(\mu) = B_c(\nu) \subseteq N'$ for all $c \in \mu(N')$, then $\mu|_{N\setminus N'} \gtrsim_f^{E'} \nu|_{N\setminus N'}$, where $E' = E|_{(N\setminus N', C\setminus C')}$.

[15] The **aggregate assignment** of N' at μ is $\mu(N') = \{c \in C | \exists i \in N' : \mu(i) = c\}$. Note that $\mu(N') = \emptyset$ if and only if all agents in N' are assigned their outside options.

3.4 Examples of Natural Stability Comparisons

Below, we present some natural stability comparisons satisfying the three properties. Some of these comparison methods are inclusion methods whereas others are the (corresponding) cardinal methods. It is easy to show that each of these stability comparisons satisfies all three properties. We will explain this only for the "blocking pairs inclusion" comparison and the "blocking pairs cardinality" comparison for illustrative purposes.

3.4.1 Blocking Pairs

The blocking pairs inclusion comparison (*pincl*) is defined as follows. For each problem $E \in \mathcal{E}$ and $\mu, \nu \in \mathcal{A}(E)$,

$$\mu \succsim^E_{pincl} \nu \Leftrightarrow B(\mu) \subseteq B(\nu).$$

That is, an assignment is *pincl*-more stable than another assignment if the set of blocking pairs in the former assignment is a subset of the set of blocking pairs in the latter assignment.[16]

The blocking pairs inclusion comparison satisfies *stability-preferred* because the set of blocking pairs for any stable assignment, which is the empty set, is trivially a subset of any other set of blocking pairs.

The *pincl* comparison satisfies *separability* because if μ is more stable than ν, then there cannot be a set of agents N' with $\mu(N') = \nu(N') = C'$ such that no agent in N' is involved in a blocking pair at ν and also no object in C' is involved in a blocking pair at ν while an agent in N' and an object in C' constitute a blocking pair at μ.

The *pincl* comparison satisfies *consistency* because if μ is more stable than ν and if for a set of agents μ and ν assignments coincide and these agents and their assigned objects are included in blocking pairs only among themselves at both μ and ν, then the set of blocking pairs at μ is still a subset of the set of blocking pairs at ν when restricted to the other agents and objects.

The blocking pairs cardinality comparison (*pcard*) is defined as follows. For each problem $E \in \mathcal{E}$ and $\mu, \nu \in \mathcal{A}(E)$,

$$\mu \succsim^E_{pcard} \nu \Leftrightarrow |B(\mu)| \leq |B(\nu)|.$$

That is, an assignment is *pcard*-more stable than another assignment if the number of blocking pairs in the former assignment is no more than the number of blocking pairs in the latter

[16] Among others, Abdulkadiroğlu et al. (2020), **?**, Tang and Zhang (2020) study this stability comparison.

assignment.[17] Note that $\succsim^E_{pincl} \subseteq \succsim^E_{pcard}$.

The blocking pairs cardinality comparison satisfies *stability-preferred* because the number of blocking pairs for any stable assignment, which is zero, is trivially no more than the number of blocking pairs at any other assignment.

The *pcard* comparison satisfies *separability* because if μ is more stable than ν and there is a set of agents N' with $\mu(N') = \nu(N') = C'$ such that no agent in N' is involved in a blocking pair at ν and also no object in C' is involved in a blocking pair at ν while an agent in N' and an object in C' constitute a blocking pair at μ, then μ must have strictly less blocking pairs than ν when restricted to the other agents and objects since overall μ has no more blocking pairs than ν.

The *pcard* comparison satisfies *consistency* because if μ is more stable than ν and if for a set of agents μ and ν assignments coincide and these agents and their assigned objects are included in blocking pairs only among themselves at both μ and ν, then the number of blocking pairs at μ is still no more than the number of blocking pairs at ν when restricted to the other agents and objects.

3.4.2 Blocking Triplets

The blocking triplets inclusion comparison (*tincl*) is defined as follows. Let $(i, j, c) \in T(\mu)$ if and only if $i \succ_c j$, $\mu(j) = c$, and $cP_i\mu(i)$. For each $E \in \mathcal{E}$ and $\mu, \nu \in \mathcal{A}(E)$,

$$\mu \succsim^E_{tincl} \nu \Leftrightarrow T(\mu) \subseteq T(\nu).$$

That is, an assignment is *tincl*-more stable than another assignment if the set of blocking triplets in the former assignment is a subset of the set of blocking triplets in the latter assignment.[18]

The blocking triplets cardinality comparison *tcard* is defined as follows. For each $E \in \mathcal{E}$ and $\mu, \nu \in \mathcal{A}(E)$,

$$\mu \succsim^E_{tcard} \nu \Leftrightarrow |T(\mu)| \leq |T(\nu)|.$$

That is, an assignment is *tcard*-more stable than another assignment if the number of blocking triplets in the former assignment is no more than the number of blocking triplets in the latter assignment. Note that $\succsim^E_{tincl} \subseteq \succsim^E_{tcard}$.

[17] Among others, **?** study this stability comparison.

[18] Kwon and Shorrer (2019) study this stability comparison, and in particular show that the blocking pairs inclusion comparison is independent from the blocking triplets inclusion comparison.

The blocking triplets inclusion and cardinality comparisons satisfy *stability-preferred*, *separability*, and *consistency*.

3.4.3 Blocking Agents

The blocking agents inclusion comparison (*aincl*) is defined as follows. Let $BA(\mu) = \{i \in N : B_i(\mu) \neq \emptyset\}$. For each problem $E \in \mathcal{E}$ and $\mu, \nu \in \mathcal{A}(E)$,

$$\mu \gtrsim^E_{aincl} \nu \Leftrightarrow BA(\mu) \subseteq BA(\nu).$$

That is, an assignment is *aincl*-more stable than another assignment if the set of blocking agents in the former assignment is a subset of the set of blocking agents in the latter assignment.

The blocking agents cardinality comparison (*acard*) is defined as follows. For each $E \in \mathcal{E}$ and $\mu, \nu \in \mathcal{A}(E)$,

$$\mu \gtrsim^E_{acard} \nu \Leftrightarrow |BA(\mu)| \leq |BA(\nu)|.$$

That is, an assignment is *acard*-more stable than another assignment if the number of blocking agents in the former assignment is no more than the number of blocking agents in the latter assignment. Note that $\gtrsim^E_{aincl} \subseteq \gtrsim^E_{acard}$.

The blocking agents inclusion and cardinality comparisons satisfy *stability-preferred*, *separability*, and *consistency*. These stability comparisons that are based on the set of blocking agents is natural because in applications such as school choice, stability has a fairness interpretation and having a minimal set of, or number of, agents who are treated unfairly (for at least one object) is a reasonable objective.

3.4.4 Blocking Objects

The blocking objects inclusion comparison (*oincl*) is defined as follows. Let $BO(\mu) = \{c \in C : B_c(\mu) \neq \emptyset\}$. For each problem $E \in \mathcal{E}$ and $\mu, \nu \in \mathcal{A}(E)$,

$$\mu \gtrsim^E_{aincl} \nu \Leftrightarrow BO(\mu) \subseteq BO(\nu).$$

That is, an assignment is *oincl*-more stable than another assignment if the set of blocking objects in the former assignment is a subset of the set of blocking objects in the latter assignment.

The blocking objects cardinality comparison (*ocard*) is defined as follows. For each $E \in \mathcal{E}$ and $\mu, \nu \in \mathcal{A}(E)$,

$$\mu \gtrsim_{acard}^{E} \nu \Leftrightarrow |BO(\mu)| \leq |BO(\nu)|.$$

That is, an assignment is *acard*-more stable than another assignment if the number of blocking agents in the former assignment is no more than the number of blocking agents in the latter assignment. Note that $\gtrsim_{oincl}^{E} \subseteq \gtrsim_{ocard}^{E}$.

The blocking objects inclusion and cardinality comparisons satisfy *stability-preferred*, *separability*, and *consistency*. These stability comparisons that are based on the set of blocking objects is natural because in applications such as school choice, stability has a fairness interpretation and having a minimal set of, or number of, objects that are allocated unfairly (by violating the priority of at least one agent) is a reasonable objective.

4 Results

Given a stability comparison f, we say that a mechanism φ is **f-more stable** than another mechanism φ' if $\varphi'(E) \gtrsim_{f}^{E} \varphi(E)$ for all $E \in \mathcal{E}$. A mechanism φ is **f-minimally unstable among *efficient* and *strategy-proof* mechanisms** if there is no other mechanism $\varphi' \neq \varphi$ that is *efficient*, *strategy-proof*, and f*-more stable* than φ.

A mechanism φ is **weakly f-minimally unstable among *efficient* and *strategy-proof* mechanisms** if there is no *efficient* and *strategy-proof* mechanism φ' that is f*-more stable* than φ and $\varphi'(E) \not\gtrsim_{f}^{E} \varphi(E)$ for some $E \in \mathcal{E}$.[19]

A mechanism φ is **robustly minimally unstable among *efficient* and *strategy-proof* mechanisms** if for any stability comparison f satisfying *stability-preferred*, *separability* and *consistency*, φ is f-minimally unstable among *efficient* and *strategy-proof* mechanisms.

The **top trading cycles (TTC) mechanism** (Abdulkadiroğlu and Sönmez, 2003) is based on Gale's TTC algorithm (Shapley and Scarf, 1974) which runs, given a problem, as follows.

TTC Algorithm:[20]

[19] If φ is f-minimally unstable among *efficient* and *strategy-proof* mechanisms, then φ is weakly f-minimally unstable among *efficient* and *strategy-proof* mechanisms (but the converse does not hold as there might exist a mechanism different from φ with the identical f-stability measure). Abdulkadiroğlu et al. (2020) and Kwon and Shorrer (2019) use the weaker second definition.

[20] **?**, **?**, and **?** propose variants of TTC for the many-to-one setup. For one-to-one problems, all variants coincide.

Step 1. Assign a counter for each object which keeps track of how many copies are still available for that object. Initially set the counters equal to the capacities of the objects. Each agent points to her top-ranked object. Each object points to the agent who has the highest priority for the object. Since the number of agents and objects are finite, there is at least one cycle. (A cycle is an ordered list of distinct agents and distinct objects $(k, c_k)_{k\in\{1,\ldots,K\}}$ such that for each $k \in \{1, \ldots, K\}$, agent k points to object c_k and object c_k points to agent $k+1$ with the convention that $K+1 = 1$. Moreover, each object can be part of at most one cycle. Similarly, each agent can be part of at most one cycle. Every agent in a cycle is assigned a copy of the object she points to and is removed. The counter of each object in a cycle is reduced by one and if it reduces to zero, the object is also removed. Counters of all other objects stay the same.

Step $\boldsymbol{t \geq 2}$. Each remaining agent points to her top-ranked object among the remaining objects and each remaining object points to the agent with highest priority among the remaining agents. There is at least one cycle. Every agent in a cycle is assigned a copy of the object that she points to and is removed. The counter of each object in a cycle is reduced by one and if it reduces to zero the object is also removed. Counters of all other objects stay the same.

Our main result is the following.

Theorem 1 *The TTC mechanism is robustly minimally unstable among efficient and strategy-proof mechanisms when objects have unit capacities.*

The proof is in Appendix A. Here, we provide a sketch of the proof and highlight the main innovative idea in the proof. Take any stability comparison f satisfying the three properties. We start with the following observation: if TTC is not f-minimally unstable among *efficient* and *strategy-proof* mechanisms, then there must be a *smallest* number of agents, say n, such that there exists a mechanism φ, different from TTC, that is defined on the domain of problems including exactly n agents and that is *strategy-proof*, *efficient*, and *f-more stable* than TTC. Note that $n \geq 3$ since at any problem including one or two agents, any mechanism that is *efficient* and *f-more stable* than TTC must coincide with TTC because TTC chooses the unique *efficient* and *stable* assignment and f satisfies *stability-preferred*.

Suppose that φ is an arbitrary mechanism that is *strategy-proof*, *efficient*, and *f-more stable than TTC* on the domain of problems with an arbitrary set of agents N where $|N| = n$ and an arbitrary set of objects C. We prove that φ coincides with TTC on this domain of problems that include (N, C). Since φ, N, and C are arbitrarily chosen, this contradicts that on the domain of problems with n agents, there exists a mechanism different from TTC that is *strategy-proof*, *efficient*, and f*-more stable* than TTC, and therefore concludes the proof.

When proving that φ coincides with TTC on this domain, the first key step is to show that φ must satisfy a *mutual-best property* (Lemma 2): If $i \in N$ and $c \in C$ are such that c is top-ranked at i's preference ordering and i has the highest priority at c, then c must be assigned to i by φ as well. If f is a set-inclusion type comparison, such as the blocking pairs inclusion comparison, then the mutual-best property is almost trivial: suppose otherwise, i.e., suppose that c is not assigned to i; then, (i, c) constitutes a blocking pair under φ, while it is not a blocking pair under TTC, contradicting that φ is more stable than TTC with respect to the blocking pairs inclusion comparison. Note that such a conclusion is not immediate if f is a cardinal type comparison, such as the blocking pairs cardinality comparison, because not assigning c to i unlike TTC does not immediately imply that the φ assignment includes more blocking pairs than the TTC assignment. Instead, we prove this by constructing a new domain of problems including fewer agents than n and constructing a new mechanism φ' that is *strategy-proof*, *efficient*, and f*-more stable than TTC* on this new domain of problems including fewer than n agents, which contradicts that n is the smallest number of agents such a domain entails.

The *mutual-best property*, together with *efficiency*, imply that the φ and TTC assignments of the agents who are assigned at the first step of the TTC algorithm must coincide. The second key step is to show that φ must satisfy a *mutual-best property with respect to any further step of the TTC algorithm* (Lemma 4): If $i \in N$ and $c \in C$ are such that i top-ranks c and c points to i at some step after the first step of the TTC algorithm, that is, they become mutually-best at some step after the first step of the TTC algorithm at E, then c must be assigned to i by φ as well. The rest of the proof shows, by induction on the step number in the TTC algorithm, that for each step of the TTC algorithm, the φ and TTC assignments of the agents who are assigned at that step of the TTC algorithm must coincide. All details are in Appendix A.

We obtain the following corollaries for several different natural stability comparison methods, some of which have been shown in the recent literature.

Corollary 1 *TTC is f-minimally unstable among efficient and strategy-proof mechanisms when objects have unit capacities if f is the*

i. blocking pairs inclusion comparison[21] *or blocking pairs cardinality comparison, or*

[21] Theorem 1 of Abdulkadiroğlu et al. (2020) show this result for the weak minimal instability among

ii. blocking triplets inclusion comparison[22] *or blocking triplets cardinality comparison, or*

iii. blocking agents inclusion comparison or blocking agents cardinality comparison, or

iv. blocking objects inclusion comparison or blocking objects cardinality comparison.

4.1 A Class where TTC is the Unique Robustly Minimally Unstable Mechanism

We will show that for a fairly large class of mechanisms, the TTC mechanism is the unique robustly minimally unstable mechanism among *efficient* and *strategy-proof* mechanisms, although Theorem 1 does not imply that TTC is the unique such mechanism in general. First, we introduce an auxiliary notion. A mechanism φ is **trivially unstable** if it chooses an unstable assignment for a unit-capacity problem where two agents find only one, and the same, object acceptable, and all other agents find no object acceptable, i.e., there exists a unit-capacity problem $E = (N, C, R, q, \succeq)$, a pair of agents $i, j \in N$, and an object $c \in C$ such that i and j find only c acceptable, each other agent finds no object acceptable, $i \succ_c j$, and $\varphi_i(E) \neq c$. Note that although TTC is not a stable mechanism, it is not trivially unstable.

Lemma 1 *If a mechanism is trivially unstable, then it is not f-minimally unstable among efficient and strategy-proof mechanisms for any stability comparison f that satisfies stability-preferred.*

Proof. Suppose that φ is *efficient*, *strategy-proof*, and *trivially unstable.* Then, there exists a unit capacity problem $E = (N, C, R, q, \succeq)$, a pair of agents $i, j \in N$, and an object $c \in C$ such that i and j find only c acceptable, each other agent finds no object acceptable, $i \succ_c j$, and $\varphi_j(E) = c$.

Let φ' be defined as follows. For each unit-capacity problem $E' = (N, C, R', q, \succeq)$ such that E and E' coincide besides the preferences of i and j (i.e., $R'_k = R_k$ for each $k \in N \setminus \{i, j\}$), let $\varphi'(E') = TTC(E')$. For every other problem E'', let $\varphi'(E'') = \varphi(E'')$. Observe that φ' is *efficient* and *strategy-proof.*

Note that on the domain of problems where φ and φ' differ, φ' is stable. Moreover, there is at least one problem, the problem E, where φ' is stable but φ is not stable. Hence, φ is

efficient and *strategy-proof* mechanisms definition.

[22] Proposition 7 of Kwon and Shorrer (2019) show this result for the weak minimal instability among *efficient* and *strategy-proof* mechanisms definition.

not f-minimally unstable among *efficient* and *strategy-proof* mechanisms for any stability comparison f that satisfies stability-preferred. $\blacksquare$

Now, consider the following class of mechanisms, called the class of **generalized TTC mechanisms**. Let h be an arbitrary function that maps each priority profile $\succeq$ to another priority profile $h(\succeq)$ consisting of the same agents and the same objects with $\succeq$. Let TTC^h be the mechanism such that for each problem $(N, C, R, q, \succeq)$, $TTC^h(N, C, R, q, \succeq) = TTC(N, C, R, q, h(\succeq))$. That is, the TTC^h outcome at each problem is obtained by running the TTC algorithm, but under the priority profile $h(\succeq)$ which may be different from the true priority profile $\succeq$. Generalized TTC mechanisms (where each member is induced by a different function) includes the serial dictatorship mechanisms where the same priority profile is used at each problem, and all objects have the same priority ordering at this common priority profile. We next show that the TTC mechanism, the one that always operates based on the true priority profile, is the unique mechanism in this class that is robustly minimally unstable among *efficient* and *strategy-proof* mechanisms when each object has unit capacity.[23]

Proposition 1 *The TTC mechanism is the unique mechanism in the class of generalized TTC mechanisms that is robustly minimally unstable among efficient and strategy-proof mechanisms when objects have unit capacities.*

Proof. Suppose that $h(\succeq) \neq \succeq$ for some priority profile $\succeq$. We will show that TTC^h is not f-minimally unstable among *efficient* and *strategy-proof* mechanisms for any stability comparison f that satisfies stability-preferred. Let $\succeq' = h(\succeq)$. Let (N, C) be the set of agents and objects included in $\succeq$. Note that there exist $c \in C$ and $i, j \in N$ such that $i \succ_c j$ and $j \succ'_c i$. Then, there exists a unit-capacity problem $E = (N, C, R, q, \succeq)$ such that i and j find only object c acceptable, each other agent finds no object acceptable, and $TTC^h_j(E) = c$. Thus, TTC^h is trivially unstable. Hence, by Lemma 1, TTC^h is not f-minimally unstable among *efficient* and *strategy-proof* mechanisms for any f that satisfies *stability-preferred* when object have unit capacities. $\blacksquare$

4.2 Non-Unit Capacities

When each object has a capacity of at least two, we characterize the capacity-priority structures for which the TTC mechanism is robustly minimally unstable among *efficient*

[23] Abdulkadiroğlu et al. (2020) show a similar result for a smaller class of mechanisms where the h function is required to satisfy the following property: for any $\succeq$, $\succeq'$, and c, if $\succeq_c = \succeq'_c$, then $\succeq^*_c = \succeq^{**}_c$ where $\succeq^* = h(\succeq)$ and $\succeq^{**} = h(\succeq')$. That is, the "distorted" priority ordering of an object only depends on the true priority ordering of the object, but not on the rest of the true priority profile.

and *strategy-proof* mechanisms. Given a set of agents N and a set of objects C, let us call $(q, \succeq)$ a **capacity-priority structure** for (N, C), where $q = (q_c)_{c \in C}$ is a capacity profile and $\succeq = (\succeq_c)_{c \in C}$ is a profile of priority orderings over N. Let $TTC^{(q,\succeq)}$ denote the TTC mechanism restricted to the domain of problems with the capacity-priority structure $(q, \succeq)$.

For each $c \in C$ and $i \in N$, let $U_c(i) = \{j \in N \setminus \{i\} : j \succ_c i\}$. A triple of agents $(i, j, k) \in N$ and a pair of objects $(c_1, c_2) \in C$ constitute a **Kesten-cycle** (**?**) if

(a) $i \succ_{c_1} j \succ_{c_1} k$ and $k \succ_{c_2} \{i, j\}$, and

(b) there is a (possibly empty) set $N_{c_1} \subseteq N \setminus \{i, j, k\}$ such that $N_{c_1} \subseteq U_{c_1}(i) \cup (U_{c_1}(j) \setminus U_{c_2}(k))$ and $|N_{c_1}| = q_{c_1} - 1$.

? showed, among other things, that $TTC^{(q,\succeq)}$ is stable (for any preference profile) if and only if $(q, \succeq)$ does not include a Kesten-cycle. Let us call $(q, \succeq)$ **acyclic** if it does not include a Kesten cycle.

Another well-known mechanism, the deferred acceptance (DA) mechanism due to Gale and Shapley (1962), will be useful in proving our next result. The **DA mechanism** associates each problem E with the assignment determined by the following algorithm.

DA Algorithm:

Step 1. Each agent proposes to her top-ranked acceptable object. If there is no such object, then she is assigned to her outside option. Each object c considers the set of proposals that it receives. Among them, it tentatively accepts the highest priority agents up to its capacity and rejects the others. If there is no rejection, then stop.

Step $t \geq 2$. Each agent who is rejected at Step $t-1$ proposes to her top-ranked acceptable object among the ones that have not rejected her yet. If there is no such object, then she is assigned to her outside option. Each object c considers the set of agents that it tentatively accepted at Step $t-1$ together with agents that have proposed at Step t. Among them, it tentatively accepts the highest priority agents up to its capacity and rejects the others. If there is no rejection, then stop. Otherwise, move to Step $t+1$.

The DA algorithm stops in finitely many steps and the DA assignment, which we denote by $DA(E)$, is defined by the acceptances at the last step. At each problem, the DA assignment is *stable* but not necessarily *efficient* (Abdulkadiroğlu and Sönmez, 2003).[24]

[24] Ergin (2002) characterized the capacity-priority structures for which the DA mechanism is *efficient*.

Theorem 2 *Let $(q, \succeq)$ be a capacity-priority structure such that all objects have non-unit capacities.*[25]

i. If $(q, \succeq)$ includes a Kesten-cycle, then $TTC^{(q,\succeq)}$ is not (weakly) f-minimally unstable among efficient and strategy-proof mechanisms for any stability comparison f that satisfies stability-preferred.

ii. The mechanism $TTC^{(q,\succeq)}$ is robustly minimally unstable among efficient and strategy-proof mechanisms if and only if $(q, \succeq)$ is acyclic.

Proof. Part i: Suppose that $(q, \succeq)$ includes a cycle. Then, there exist a triple of agents $(i, j, k) \in N$ and a pair of objects $(c_1, c_2) \in C$ such that

(a) $i \succ_{c_1} j \succ_{c_1} k$ and $k \succ_{c_2} \{i, j\}$, and

(b) there is a (possibly empty) set $N_{c_1} \subseteq N \backslash \{i, j, k\}$ such that $N_{c_1} \subseteq U_{c_1}(i) \cup (U_{c_1}(j) \backslash U_{c_2}(k))$ and $|N_{c_1}| = q_{c_1} - 1$.

Let D_1 be the domain of problems with the capacity-priority structure $(q, \succeq)$ such that for each problem $E = (N, C, R, q, \succeq) \in D_1$, each agent in N_{c_1} finds only c_1 acceptable (therefore top-ranks c_1), each agent in $\{i, j, k\}$ top-ranks c_1 or c_2 (but possibly finds other objects acceptable), and every other agent in $N \setminus (N_{c_1} \cup \{i, j, k\})$ finds no object acceptable. Let D_2 be the remaining set of problems with the capacity-priority structure $(q, \succeq)$.

Let φ be defined as follows. For each $E \in D_1$, let $\varphi(E) = DA(E)$; and for each $E \in D_2$, let $\varphi(E) = TTC(E)$.

Observe that φ is *efficient* and *stable* on the domain D_1, since each agent in N_{c_1} is always assigned to c_1 (because they are in the top-q_{c_1} priority class at $\succeq_{c_1}$ among agents who find at least one object acceptable) and there is never a rejection cycle in the DA algorithm since $q_{c_2} > 1$.

We claim that φ is *strategy-proof*. Consider any agent $s \in N_{c_1}$. Agent s does not have a profitable manipulation at any problem in D_1 since s receives his top choice. Agent s does not have a profitable manipulation at any problem in D_2 neither since by misreporting his preferences, he either induces a problem also in D_2 and such a manipulation is not profitable by the *strategy-proofness* of TTC, or he induces a problem in D_1 and receives c_1 which cannot

[25] Requiring all objects to have non-unit capacities, as opposed to only some of the objects, ensures that in a restricted domain that we identify in the proof, the DA mechanism is always efficient, which plays an important role in the proof.

be better than what he receives by his truthful report under TTC since he belongs to the top-q_{c_1} priority class at $\succeq_{c_1}$ among agents who find at least one object acceptable.

Consider any agent $s \in \{i, j, k\}$. Suppose that s has a profitable manipulation at E. Then $\varphi_s(E)$ cannot be s's top ranked object under E.

Suppose that $E \in D_1$. By *strategy-proofness* of DA, it must be that by profitably misreporting his preferences, s induces a problem $E' \in D_2$. Let c be the object s top-ranks at E'. Note that $c \notin \{c_1, c_2\}$.

Case 1: $\varphi_s(E') = c$. Since this is a profitable manipulation, $\varphi_s(E)$ is worse than c for s at E. By $q_c \geq 2$, the agents in $\{i, j, k\} \backslash \{s\}$ must receive c at $\varphi(E)$. But then s must receive his top-ranked object under $\varphi(E)$, a contradiction.

Case 2: $\varphi_s(E') \neq c$. Then, by $q_c \geq 2$, it must be that the agents in $\{i, j, k\} \backslash \{s\}$ must receive c at $\varphi(E')$. But then either s receives c_1 and at least one agent in $\{i, j, k\} \backslash \{s\}$ top-ranks c_1, or the top-ranked object of at least one agent in $\{i, j, k\} \backslash \{s\}$ has an available copy under $\varphi(E')$, which are both contradictions to efficiency of $\varphi(E') = TTC(E')$.

Suppose that $E \in D_2$. By *strategy-proofness* of TTC, it must be that by profitably misreporting his preferences, s induces a problem $E' \in D_1$. Let c be the object s top-ranks at E. Note that $c \notin \{c_1, c_2\}$. Since $\varphi_s(E) \neq c$ and $q_c \geq 2$, it must be that the agents in $\{i, j, k\} \setminus \{s\}$ receive c at $\varphi(E)$. But then either s receives c_1 and at least one agent in $\{i, j, k\} \backslash \{s\}$ top-ranks c_1, or the top-ranked object of at least one agent in $\{i, j, k\} \backslash \{s\}$ has an available copy under $\varphi(E)$ (because $E' \in D_1$), which are both contradictions to efficiency of $\varphi(E) = TTC(E)$.

It is also easy to see that no agent in $N \setminus (N_{c_1} \cup \{i, j, k\})$ has a profitable manipulation at any problem. Hence, φ is *strategy-proof*.

Finally, note that on the domain of problems where φ and TTC differ, φ is stable. Moreover, for the following problem, φ is stable but TTC is not stable: each agent in N_{c_1} find only c_1 acceptable, i finds only c_2 acceptable, j and k find only c_1 acceptable. Hence, TTC is not (weakly) f-minimally unstable among *efficient* and *strategy-proof* mechanisms for any stability comparison f satisfying stability-preferred.

Part ii: If $(q, \succeq)$ is acyclic, then $TTC^{(q,\succeq)}$ is stable at each problem (**?**), and therefore $TTC^{(q,\succeq)}$ is robustly minimally unstable among *efficient* and *strategy-proof* mechanisms. The other direction follows from Part i. ■

When objects may have multiple available copies, Abdulkadiroğlu et al. (2020) provide an example of a capacity-priority structure $(q, \succeq)$ such that the $DA^{(q,\succeq)}$ is *efficient* and *stable* at every problem, while $(q, \succeq)$ is not *acyclic* (in the sense of **?**) and therefore $TTC^{(q,\succeq)}$ is not *stable* at every problem. Based on this, Abdulkadiroğlu et al. (2020) show that there exists a *strategy-proof* and *efficient* mechanism φ that is *more stable* than TTC (on the full domain): simply consider the mechanism φ that coincides with DA on the domain of problems including $(q, \succeq)$, and coincides with TTC elsewhere. Our Theorem 2 reveals that the failure of the minimal instability of TTC for a given capacity-priority structure when we allow for multiple available copies per object is not because the DA mechanism is always *efficient* while the TTC mechanism is sometimes *unstable*, but it is solely because the TTC mechanism is sometimes unstable for the given capacity-priority profile.[26]

Remark 2 *When objects may have weak priorities (i.e., when indifference among different agents is allowed), the standard approach is to convert the weak priorities into strict priorities using a predetermined tie-breaker rule and then run the TTC mechanism. With weak priorities, even when objects have unit capacities, Theorem 1 does not extend for a similar reason as in the multi-capacity case. Given a tie-breaker rule, it is possible to show that the TTC mechanism is not minimally unstable among efficient and strategy-proof mechanisms, by formalizing the following intuition: when the tie-breaker is applied, a priority profile involving a Kesten-cycle may occur, although with another tie-breaker Kesten-cycles could be avoided. This was shown in an earlier working paper version of Abdulkadiroğlu et al. (2020).*

5 Conclusion

We have formulated natural methods to compare assignments in terms of their stability in the context of priority-based resource allocation. We have shown that the TTC mechanism is minimally unstable among *efficient* and *strategy-proof* mechanisms in a robust sense—that is, for any natural stability comparison—when each object has unit capacity. This is a strong justification for using the TTC mechanism in applications where objects have unit capacities. When objects have non-unit capacities, we have characterized the capacity-priority structures for which this justification is preserved, which turns out to be a very limited set of capacity-priority structures. Overall, our paper sheds further light on how the TTC mechanism incorporates priorities, contributing to the theoretical literature on understanding popular resource allocation mechanisms and to the policy debates on which mechanism to

[26] For non-unit capacities, it is an open question whether for the variants of TTC proposed by **?**, **?**, and **?** a parallel result to Theorem 2 holds. For this, first one would have to characterize the capacity-priority-structures for which any of these variants of TTC is stable (for any preference profile). This is only known for the TTC algorithm used here.

use in practice.

References

Abdulkadiroğlu, Atila and Tayfun Sönmez, "School choice: A mechanism design approach," *American Economic Review*, June 2003, *93* (3), 729–747.

_ **, Yeon-Koo Che, Parag A. Pathak, Alvin E. Roth, and Olivier Tercieux**, "Efficiency, Justified Envy, and Incentives in Priority-Based Matching," *American Economic Review: Insights, forthcoming*, 2020.

Abraham, J., P. Biró, and D. F. Manlove, "Almost stable matchings in the Roommates problem," in "Proceedings of WAOA 2005: the 3rd workshop on approximation and online algorithms," Vol. 3879, Springer, 2005, pp. 1–14.

Bonkoungou, Somouaoga and Alexander Nesterov, "Reforms Meet Fairness Concerns In School And College Admissions," *Working paper, available at SSRN: https://ssrn.com/abstract=3664089 or http://dx.doi.org/10.2139/ssrn.3664089*, 2020.

Ergin, Haluk I., "Efficient resource allocation on the basis of priorities," *Econometrica*, 2002, *70* (6), 2489–2497.

Gale, David and Lloyd S. Shapley, "College Admissions and the Stability of Marriage," *The American Mathematical Monthly*, jan 1962, *69* (1), 9–15.

Kesten, Onur, "School choice with consent," *The Quarterly Journal of Economics*, 2010, *125* (3), 1297–1348.

_ **and Morimitsu Kurino**, "Strategy-proof improvements upon deferred acceptance: A maximal domain for possibility," *Games and Economic Behavior*, 2019, *117*, 120–143.

Kwon, Hyukjun and Ran I. Shorrer, "Justified-Envy Minimal Mechanisms in School Choice," *Working Paper Available at SSRN: https://ssrn.com/abstract=3495266*, 2019.

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

Pápai, Szilvia, "Strategyproof Assignment by Hierarchical Exchange," *Econometrica*, 2000, *68*, 1403–1433.

Pycia, Marek and Utku Ünver, "Incentive Compatible Allocation and Exchange of Discrete Resources," *Theoretical Economics*, 2017, *12*, 287–329.

Romm, Assaf, Alvin E. Roth, and Ran I. Shorrer, "Stability vs. No Justified Envy," *Working Paper Available at SSRN: https://ssrn.com/abstract=3550122*, 2020.

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

Shapley, Lloyd and Herbert Scarf, "On cores and indivisibility," *Journal of Mathematical Economics*, 1974, *1* (1), 23 – 37.

Svensson, Lars-Gunnar, "Strategy-proof allocation of indivisible goods," *Social Choice and Welfare*, 1999, *16*, 557–567.

Tang, Qianfeng and Yongchao Zhang, "Weak Stability and Pareto Efficiency in School Choice," *Economic Theory, forthcoming*, 2020.

Appendix A Proof of Theorem 1

Suppose not. Then, there exists a *smallest* number of agents, say n, such that there exists a mechanism, different from TTC, that is defined on the domain of problems including *exactly* n agents and that is *strategy-proof*, *efficient*, and *f-more stable* than TTC for a stability comparison f satisfying *stability-preferred*, *separability* and *consistency*. Note that $n \geq 3$ since at any problem including 1 or 2 agents, any mechanism that is *efficient* and f-more stable than TTC must coincide with TTC because TTC chooses the unique *efficient* and *stable* assignment and f satisfies *stability-preferred.*

Suppose that φ is an arbitrary mechanism that is *strategy-proof*, *efficient*, and *f-more stable than TTC* on the domain of problems with an arbitrary set of agents N where $|N| = n$ and an arbitrary set of objects C with unit capacities. We prove that φ coincides with TTC on this domain of problems that include (N, C). Since φ, N, and C are arbitrarily chosen, this contradicts that on the domain of problems with n agents, there exists a mechanism different from TTC that is *strategy-proof*, *efficient*, and *f-more stable* than TTC, and therefore concludes the proof. In what follows, let D denote the domain of unit-capacity problems that have N as the set of agents and C as the set of objects.

We will sometimes denote a problem simply by its preference profile, i.e., $\varphi(R)$ instead of $\varphi(E)$, when the rest of the problem in question is clear. Also, when we write $R_i : cc'$, we mean cR_ic' and that any other object is unacceptable.

Lemma 2 *Let $E = (N, C, R, q, \succeq) \in D$. Let $i \in N$ and $c \in C$ be such that c is top-ranked at i's preference ordering and i has the highest priority at c, that is, they are mutually-best at the first step of the TTC algorithm at E. Then, $\varphi_i(R) = c$.*

Proof. Suppose not, i.e., suppose that $\varphi_i(R) \neq c$. Let R'_i be a preference ordering for i at which c is the only acceptable object. By *strategy-proofness*, $\varphi_i(R'_i, R_{-i}) = \emptyset$. By *efficiency*, there exists $j_1 \neq i$ such that $\varphi_{j_1}(R'_i, R_{-i}) = c$. Let R'_{j_1} be a preference relation for agent j_1 at which c is the only acceptable object. By *strategy-proofness*, $\varphi_{j_1}(R'_i, R'_{j_1}, R_{-\{i,j_1\}}) = c$.

Now, suppose that there exists a preference profile $\overline{R}_{-\{i,j_1\}}$ of agents $N \setminus \{i, j_1\}$ such that $\varphi_c(R'_i, R'_{j_1}, \overline{R}_{-\{i,j_1\}}) \in N \setminus \{i, j_1\}$, i.e., c is assigned to an agent different from i or j_1. Let $j_2 \in N \setminus \{i, j_1\}$ such that $\varphi_{j_2}(R'_i, R'_{j_1}, \overline{R}_{-\{i,j_1\}}) = c$. Let R'_{j_2} be a preference relation for agent j_2 at which c is the only acceptable object. By *strategy-proofness*, $\varphi_{j_2}(R'_i, R'_{j_1}, R'_{j_2}, \overline{R}_{-\{i,j_1,j_2\}}) = c$.

Successive applications of the above argument imply that there exist $\{j_1, \ldots, j_m\}$ and a preference profile $R^*_{-\{i,j_1,j_2,\ldots,j_m\}}$ for agents $N \setminus \{i, j_1, j_2, \ldots, j_m\}$ such that

- for each $t \in \{1, \ldots, m\}$, R'_{j_t} is a preference relation for agent j_t at which c is the only acceptable object,
- $\varphi_{j_m}(R'_i, R'_{j_1}, R'_{j_2}, \ldots, R'_{j_m}, R^*_{-\{i,j_1,j_2,\ldots,j_m\}}) = c$, and
- for any preference profile $R^{**}_{-\{i,j_1,j_2,\ldots,j_m\}}$ for agents $N \setminus \{i, j_1, j_2, \ldots, j_m\}$, we have $\varphi_c(R'_i, R'_{j_1}, R'_{j_2}, \ldots, R'_{j_m}, R^{**}_{-\{i,j_1,j_2,\ldots,j_m\}}) \in \{i, j_1, \ldots, j_m\}$.

Let $R' = (R'_i, R'_{j_1}, R'_{j_2}, \ldots, R'_{j_m}, R^*_{-\{i,j_1,j_2,\ldots,j_m\}})$. Note that if $m = n$, then $|B(TTC(R'))| = 0 < |B(\varphi(R'))|$ since i has the highest priority among all agents at c. Moreover, this contradicts that φ is *f-more stable* than TTC as f satisfies stability-preferred and $TTC(R') \gtrsim_f^{R'} \varphi(R')$. Thus, $m < n$.

Now, we will construct a mechanism φ' defined on the domain of problems with agents $N' = N \setminus \{i, j_1, j_2, \ldots, j_m\}$ and objects $C' = C \setminus \{c\}$ that is *strategy-proof*, *efficient*, and *f-more stable* than TTC, which will contradict that n is the smallest number of agents such a domain entails.

Let φ' be defined as follows. For each preference profile $\overline{R}_{N'}$ of N',

- If for each $j \in N'$, $\overline{R}_j$ agrees with R^*_j on the relative orderings of $C' \cup \{\emptyset\}$, then $\varphi'(\overline{R}_{N'}) = \varphi(R'_i, R'_{j_1}, R'_{j_2}, \ldots, R'_{j_m}, R^*_{N'})|_{N'}$
- If for each $j \in N'' \subseteq N'$, $\overline{R}_j$ agrees with R^*_j on the relative orderings of $C' \cup \{\emptyset\}$, and for each $j' \in N' \setminus N''$, $\overline{R}_j$ does not agree with $R^*_{j'}$ on the relative orderings of $C' \cup \{\emptyset\}$, then let $\varphi'(\overline{R}_{N'}) = \varphi(R'_i, R'_{j_1}, R'_{j_2}, \ldots, R'_{j_m}, R'_{N'})|_{N'}$ where for each $j \in N''$, $R'_j = R^*_j$, and for each $j \in N' \setminus N''$, R'_j is a preference ordering which bottom-ranks c and agrees with $\overline{R}_{j'}$ on the relative orderings of $C' \cup \{\emptyset\}$.

Note that φ' is well-defined, in particular when $\{i, j_1, \ldots, j_m\}$ report $(R'_i, R'_{j_1}, R'_{j_2}, \ldots, R'_{j_m})$, no agent in N' can receive object c under any preference profile of N'. To see that φ' is *strategy-proof*, observe that manipulability of φ' would immediately imply the manipulability of φ. *Efficiency* of φ' also follows directly from *efficiency* of φ. We will next show that φ' is *f-more stable* than TTC.

Note that at any problem R (in the domain where there are n agents) such that $(i, j_1, \ldots, j_m)$ report $(R'_i, R'_{j_1}, R'_{j_2}, \ldots, R'_{j_m})$, no agent in $(i, j_1, \ldots, j_m)$ is involved in a blocking pair at the TTC assignment; moreover, no agent in N' is included in a blocking pair together with c at the TTC assignment. Thus, $B(TTC_x(R)) = \emptyset$ for all $x \in (N \backslash N') \cup \{c\}$. On the other hand, consider the problem $R' = (R'_i, R'_{j_1}, R'_{j_2}, \ldots, R'_{j_m}, R^*_{-\{i, j_1, j_2, \ldots, j_m\}})$. Note that (i, c) is a blocking pair at $\varphi(R')$. Since φ is *f-more stable* than TTC, we have $\varphi(R') \gtrsim_f^{R'} TTC(R')$. Furthermore, by efficiency of φ and construction, we have $\cup_{h \in N \backslash N'} \{\varphi_h(R')\} = \{c\} = \cup_{h \in N \backslash N'} \{TTC_h(R')\}$. Consequently, by separability of f, at the problem $\overline{R}_{N'}$ (in the domain where there are $n - m$ agents) where for each $j \in N'$, $\overline{R}_j$ agrees with R^*_j on the relative orderings of C', $\varphi'(\overline{R}_{N'}) = \varphi(R')|_{N'} \gnsim_f^{\overline{R}_{N'}} TTC(R')|_{N'} = TTC(\overline{R}_{N'})$ (where the equalities follow from the definition of φ' and TTC).

Now consider any problem R (in the domain where there are n agents) such that $(i, j_1, \ldots, j_m)$ report $(R'_i, R'_{j_1}, R'_{j_2}, \ldots, R'_{j_m})$. Then $B_x(TTC(R)) = \emptyset$ for all $x \in (N \backslash N') \cup \{c\}$. If for some $i \in N \backslash N'$, $\varphi_i(R) \neq \emptyset$, then by efficiency $\varphi_i(R) = c$ and we use the same arguments as above. If $TTC_i(R) = \varphi_i(R)$ for all $i \in N \backslash N'$, then by construction, $B_x(TTC(R')) = B_x(\varphi(R'))$ for all $x \in N \backslash N'$. Hence, by consistency of f and $\varphi(R) \gtrsim_f^R TTC(R)$, we obtain $\varphi(R)|_{N'} \gtrsim_f^{R_{N'}} TTC(R)|_{N'}$. Thus (as R was arbitrary), for any profile $\overline{R}_{N'}$ of N' we have $\varphi'(\overline{R}_{N'}) \gtrsim_f^{R_{N'}} TTC(\overline{R}_{N'})$ (from the definition of φ' and TTC). Hence, φ' is f-more stable than TTC, contradicting that n is the smallest number of agents such a domain entails. ■

Lemma 3 *Let $E = (N, C, R, q, \succeq) \in D$. Let $i \in N$ be an agent who is assigned an object at Step* 1 *of $TTC(R)$. Then, $\varphi_i(R) = TTC_i(R)$.*

Proof. Let I_1 denote the set of agents who are assigned an object at Step 1 of $TTC(R)$ and C_1 denote the set of objects that are allocated at Step 1 of $TTC(R)$. Note that if for each $i \in I_1$, $\varphi_i(R) \in C_1$, then by *efficiency*, $\varphi_i(R) = TTC_i(R)$ for each $i \in I_1$.

Suppose that there exists $i_1 \in I_1$ such that $\varphi_{i_1}(R) \notin C_1$. Let $c_1 \in C_1$ be the object that is assigned at Step 1 of $TTC(R)$ and points to i_1 in Step 1 of $TTC(R)$. Let R'_{i_1} be a preference ordering for i_1 at which $TTC_{i_1}(R)$ is top-ranked and c_1 is second-ranked, i.e. $R'_{i_1} : TTC_{i_1}(R)c_1$. By strategy-proofness, $\varphi_{i_1}(R'_{i_1}, R_{-i_1}) \neq TTC_{i_1}(R)$. By Lemma 2 and *strategy-proofness*, $\varphi_{i_1}(R'_{i_1}, R_{-i_1}) = c_1$.

Note that I_1 is still the set of agents who are assigned an object at Step 1 of $TTC(R'_{i_1}, R_{-i_1})$ and C_1 is still the set of objects that are allocated at Step 1 of $TTC(R'_{i_1}, R_{-i_1})$. Now, if for each $i \in I_1 \setminus \{i_1\}$, $\varphi_i(R) \in C_1$, then *efficiency* would imply that $\varphi_i(R'_{i_1}, R_{-i_1}) = TTC_i(R'_{i_1}, R_{-i_1})$ for each $i \in I_1$, which would contradict $\varphi_{i_1}(R'_{i_1}, R_{-i_1}) \neq TTC_{i_1}(R) = TTC_{i_1}(R'_{i_1}, R_{-i_1})$. Thus, there exists $i_2 \in I_1 \setminus \{i_1\}$ such that $\varphi_{i_2}(R) \notin C_1$. Let $c_2 \in C_1$ be the object that is assigned in Step 1 of $TTC(R)$ and points to i_2 in Step 1 of $TTC(R)$. Let R'_{i_2} be a preference ordering for i_2 at which $TTC_{i_2}(R)$ is top-ranked and c_2 is second-ranked. By strategy-proofness, $\varphi_{i_2}(R'_{i_1}, R'_{i_2}, R_{-\{i_1,i_2\}}) \neq TTC_{i_2}(R)$. By Lemma 2 and *strategy-proofness*, $\varphi_{i_2}(R'_{i_1}, R'_{i_2}, R_{-\{i_1,i_2\}}) = c_2$.

Continuing in a similar fashion, we identify a list of agents $(i_1, \ldots, i_m)$ and a preference profile $R' = (R'_{i_1}, \ldots, R'_{i_m}, R_{-\{i_1,\ldots,i_m\}})$ such that $\{i_1, \ldots, i_m\} \subseteq I_1$, $\varphi_i(R') \in C_1$ for each $i \in I_1$, and $\varphi_{i_m}(R') \neq TTC_{i_m}(R')$, which contradicts *efficiency* of φ. ■

Lemma 4 *Let k be a number. Suppose that at any problem $\overline{E} = (N, C, \overline{R}, q, \succeq) \in D$, if an agent i is assigned an object at an earlier step than Step k at $TTC(\overline{R})$, then $\varphi_i(\overline{R}) = TTC_i(\overline{R})$. Let $E = (N, C, R, q, \succeq) \in D$. Suppose that $i \in N$ and $c \in C$ are such that i top-ranks c and c points to i at Step k of the TTC algorithm at E, that is, they become mutually-best at Step k of the TTC algorithm at E. Then, $\varphi_i(R) = c$.*

Proof. Note that the statement is true for $k = 1$ by Lemma 2. We show the statement for $k > 1$. Suppose not, i.e., suppose that $\varphi_i(R) \neq c$. Let R'_i be a preference relation for agent i at which c is the only acceptable object. By *strategy-proofness* and *efficiency*, $\varphi_i(R'_i, R_{-i}) = \emptyset$. By *efficiency*, there exists $j_1 \neq i$ such that $\varphi_{j_1}(R'_i, R_{-i}) = c$. Let $I_{<k}$ denote the set of agents who are assigned copies at an earlier step than Step k at $TTC(R)$. Note that any agent $j \in I_{<k}$ is still assigned the same copy at an earlier step than Step k at $TTC(R'_i, R_{-i})$. Then, by our supposition, for any agent $j \in I_{<k}$, $\varphi_j(R) = TTC_j(R)$. But then, $j_1 \notin I_{<k}$. Hence, by the definition of TTC, i has higher priority than j_1 at c since c points to i at Step k of

$TTC(R)$. Let R'_{j_1} be a preference relation for agent j_1 at which c is the only acceptable object. By *strategy-proofness*, $\varphi_{j_1}(R'_i, R'_{j_1}, R_{-\{i,j_1\}}) = c$.

Now, suppose that there exists a preference profile $\overline{R}_{-\{i,j_1,I_{<k}\}}$ of agents $N \setminus (\{i, j_1\} \cup I_{<k})$ such that $\varphi_c(R'_i, R'_{j_1}, R_{I_{<k}}, \overline{R}_{-\{i,j_1,I_{<k}\}}) \in N\setminus(\{i, j_1\}\cup I_{<k})$. Let $j_2 \in N\setminus(\{i, j_1\}\cup I_{<k})$ be such that $\varphi_{j_2}(R'_i, R'_{j_1}, R_{I_{<k}}, \overline{R}_{-\{i,j_1,I_{<k}\}}) = c$. Let R'_{j_2} be a preference relation for agent j_2 at which c is the only acceptable object. By *strategy-proofness*, $\varphi_{j_2}(R'_i, R'_{j_1}, R'_{j_2}, R_{I_{<k}}, \overline{R}_{-\{i,j_1,j_2,I_{<k}\}}) = c$.

Successive applications of the above argument imply that there exist $\{j_1, \ldots, j_m\}$ and a preference profile $R^*_{-\{i,j_1,j_2,\ldots,j_m,I_{<k}\}}$ for agents $N \setminus (\{i, j_1, j_2, \ldots, j_m\} \cup I_{<k})$ such that

- for each $t \in \{1, \ldots, m\}$, R'_{j_t} is a preference relation for agent j_t at which c is the only acceptable object,
- $\varphi_{j_m}(R'_i, R'_{j_1}, R'_{j_2}, \ldots, R'_{j_m}, R_{I_{<k}}, R^*_{-\{i,j_1,j_2,\ldots,j_m,I_{<k}\}}) = c$, and
- for any preference profile $R^{**}_{-\{i,j_1,j_2,\ldots,j_m,I_{<k}\}}$ for agents $N \setminus (\{i, j_1, j_2, \ldots, j_m\} \cup I_{<k})$, we have $\varphi_c(R'_i, R'_{j_1}, R'_{j_2}, \ldots, R'_{j_m}, R_{I_{<k}}, R^{**}_{-\{i,j_1,j_2,\ldots,j_m,I_{<k}\}}) \notin N \setminus (\{i, j_1, j_2, \ldots, j_m\} \cup I_{<k})$.

Let $R' = (R'_i, R'_{j_1}, R'_{j_2}, \ldots, R'_{j_m}, R_{I_{<k}}, R^*_{-\{i,j_1,j_2,\ldots,j_m,I_{<k}\}})$. First note that, each $j \in I_{<k}$ is still assigned the same object as in $TTC(R)$ at an earlier step than Step k at $TTC(R')$. Hence, by our supposition, for any $j \in I_{<k}$, $\varphi_j(R') = TTC_j(R') = TTC_j(R)$. For later purposes, let $c = c^1$ and $J_{c^1} = \{i, j_1, j_2, \ldots, j_m\}$.

First we show that $m < n - |I_{<k}|$. Suppose that $m = n - |I_{<k}|$. Note that $N = J_{c^1} \cup I_{<k}$ and $TTC_j(R') = \varphi_j(R')$ for all $j \in I_{<k}$. Let $C' = \cup_{j\in I_{<k}} TTC_j(E')$ denote the aggregate assignment of $I_{<k}$. Note that for all $j \in I_{<k}$ we have $B_j(TTC(R')) = B_j(\varphi(R')) \subseteq C'$ and for all $c \in C'$, $B_c(TTC(R')) = B_c(\varphi(R')) \subseteq I_{<k}$. But then $\varphi(R') \gtrsim_f^{R'} TTC(R')$ and consistency of f imply $\varphi(R')|_{J_{c^1}} \gtrsim_f^{E''} TTC(R')|_{J_{c^1}}$ where $E'' = E'|_{(J_{c_1}, C\setminus C')}$ (where $E' = (N, C, R', q, \succeq)$). But this is a contradiction to stability-preferred of f because at the problem E'' we have $B(TTC(R')|_{J_{c^1}}) = \emptyset \neq B(\varphi(R')|_{J_{c^1}})$ since i has the highest priority among agents in J_{c^1} at c^1. Thus, $m < n - |I_{<k}|$.

Next we show that for all $j \in I_k\setminus J_{c^1}$ we have $TTC_j(R') = \varphi_j(R')$. If $\cup_{j\in I_k\setminus J_{c^1}} TTC_j(R') = \cup_{j\in I_k\setminus J_{c^1}} \varphi_j(R')$, then this follows from efficiency of $\varphi(R')$ and $TTC(R')$. Thus, for some $j \in I_k\setminus J_{c^1}$, $\varphi_j(R') \notin \cup_{h\in I_k}\{TTC_h(R')\}$. Thus, by construction of J_{c^1} and the induction hypothesis, $\varphi_j(R') \notin \{c\} \cup [\cup_{h\in I_{\leq k}}\{TTC_h(R')\}]$ (where $I_{\leq k} = I_{<k} \cup I_k$). Let $j = h_l$ belong in $TTC(R')$ to a cycle $c_1 \to h_1 \to \cdots \to c_l \to h_l \to c_{l+1} \to h_{l+1} \to \cdots \to c_1$ but $\varphi_{h_l}(R') \neq c_{l+1}$, i.e. $TTC_{h_l}(R') = c_{l+1}$, $TTC_{h_{l-1}}(R') = c_l$ and c_l points to h_l in the TTC-algorithm. Let

$\hat{R}_{h_l} : c_{l+1}c_l$ and let $\hat{R}$ be the problem obtained from R' by only changing the preference ordering of agent h_l to $\hat{R}_{h_l}$. By strategy-proofness and efficiency, $\varphi_{h_l}(\hat{R}) = \emptyset$ or $\varphi_{h_l}(\hat{E}) = c_l$.

If $\varphi_{h_l}(\hat{R}) = c_l$, then $\varphi_{h_{l-1}}(\hat{R}) \neq c_l$. Then let $\hat{R}_{h_{l-1}} : c_l c_{l-1}$ and let $\hat{R}'$ be the problem obtained from $\hat{R}$ by only changing the preference ordering of agent h_{l-1} to $\hat{R}_{h_{l-1}}$. By strategy-proofness and efficiency, $\varphi_{h_{l-1}}(\hat{R}') = \emptyset$ or $\varphi_{h_{l-1}}(\hat{R}') = c_{l-1}$. In the latter case, again we have $\varphi_{h_{l-2}}(\hat{R}') \neq c_{l-1}$, and so on until each agent h_l receives c_l and we find a contradiction to efficiency. Thus, at some point for $h_t \in I_k \backslash J_{c^1}$ and $R_{h_t} : c_{t+1}c_t$ we have for the constructed profile R, $\varphi_{h_t}(R) = \emptyset$. Then let $R''_{h_t} : c_t$ and $R'' = (R''_{h_t}, R_{-h_t})$.

But then set $c^2 \equiv c_t$. Analogous successive applications of the above arguments show that there exists J_{c^2} and a preference profile $R''_{J_{c^2}}$ such that for all $i \in J_{c^2}$, $R''_i : c^2$, and a preference profile $R^*_{-J_{c^1} \cup J_{c^2} \cup I_{<k}}$ for agents $N \setminus (J_{c^1} \cup J_{c^2} \cup I_{<k})$ such that

- for each $i \in J_{c^2}$, R''_i is a preference relation for agent i at which c^2 is the only acceptable object,
- $\varphi_h(R'_{J_{c^1}}, R''_{J_{c^2}}, R_{I_{<k}}, R^*_{-J_{c^1} \cup J_{c^2} \cup I_{<k}}) = c^1$ for some $h \in J_{c^1}$,
- $\varphi_h(R'_{J_{c^1}}, R''_{J_{c^2}}, R_{I_{<k}}, R^*_{-J_{c^1} \cup J_{c^2} \cup I_{<k}}) = c^2 \neq TTC_h(R'_{J_{c^1}}, R''_{J_{c^2}}, R_{I_{<k}}, R^*_{-J_{c^1} \cup J_{c^2} \cup I_{<k}})$ for some $h \in J_{c^2}$, and
- for any preference profile $R^{**}_{-J_{c^1} \cup J_{c^2} \cup I_{<k}}$ for agents $N \setminus (J_{c^1} \cup J_{c^2} \cup I_{<k})$, we have $\varphi_{c^2}(R'_{J_{c^1}}, R''_{J_{c^2}}, R_{I_{<k}}, R^{**}_{-J_{c^1} \cup J_{c^2} \cup I_{<k}}) \notin N \setminus (J_{c^1} \cup J_{c^2} \cup I_{<k})$.

Let $R'' = (R'_{J_{c^1}}, R''_{J_{c^2}}, R_{I_{<k}}, R^*_{-J_{c^1} \cup J_{c^2} \cup I_{<k}})$. If for some profile $R = (R'_{J_{c^1}}, R''_{J_{c^2}}, R_{I_{<k}}, R^{**}_{-J_{c^1} \cup J_{c^2} \cup I_{<k}})$ and some $j \in I_k \backslash (J_{c^1} \cup J_{c^2})$ we have $\varphi_j(R) \neq TTC_j(R)$, then we do the same as above and find c^3 and J_{c^3} together with a profile $R'''_{J_{c^3}}$ (and continue).

Otherwise we have for any problem $R = (R'_{J_{c^1}}, R''_{J_{c^2}}, R_{I_{<k}}, R^{**}_{-J_{c^1} \cup J_{c^2} \cup I_{<k}})$ and all $j \in I_k \backslash (J_{c^1} \cup J_{c^2})$, $\varphi_j(R) = TTC_j(R)$.

Now consider R'' and I_{k+1}. If for some $j \in I_{k+1} \backslash (J_{c^1} \cup J_{c^2})$, $\varphi_j(R) \neq TTC_j(R)$, then we find as above c^3 and J_{c^3}, and so on.

Thus, we find $\{c^1, \ldots, c^q\}$ and mutually disjunct sets $J_{c^1}, \ldots, J_{c^q}$ and $I_{<k}$ such that for $R^{(q)} = (R'_{J_{c^1}}, R''_{J_{c^2}}, \ldots, R^{(q)}_{J_{c^q}}, R_{I_{<k}}, R^*_{-J_{c^1} \cup J_{c^2} \cup \cdots \cup J_{c^q} \cup I_{<k}})$ and $E^{(q)} = (N, C, R^{(q)}, q, \succeq)$ we have

- for each $j \in J_{c^p}$ (with $p \in \{1, \ldots, q\}$), $R^{(p)}_j$ is a preference relation for agent j at which c^p is the only acceptable object,
- for each $p \in \{1, \ldots, q-1\}$, $\varphi_h(E^{(p)}) = c^p$ for some $h \in J_{c^p}$,

- $\varphi_h(E^{(q)}) = c^q \neq TTC_h(E^{(q)})$ for some $h \in J_{c^q}$, and
- $\varphi_j(E^{(q)}) = TTC_j(E^{(q)})$ for all $j \in N\setminus(J_{c^1} \cup \cdots \cup J_{c^q})$.

Let $\mu = \varphi(E^{(q)})$ and $\nu = TTC(E^{(q)})$. Because φ is f-more stable than TTC, we have $\mu \succsim_f^{E^{(q)}} \nu$. Now we will successively remove in the order $J_{c^1}, \ldots, J_{c^q}$.

If $\mu(j) = \nu(j)$ for all $j \in J_{c^1}$, then we have $B_j(\mu) = B_j(\nu) = \emptyset$ for all $j \in J_{c^1}$ and $B_{c^1}(\mu) = \emptyset = B_{c^1}(\nu)$. Thus, by *consistency* of f we obtain $\mu|_{N\setminus J_{c^1}} \succsim_f^{E^1} \nu|_{N\setminus J_{c^1}}$ where $E^1 = E^{(q)}|_{(N\setminus J_{c_1}, C\setminus\{c^1\})}$. Otherwise, $B_j(\nu) = \emptyset$ for all $j \in J_{c^1}$ and $B_{c^1}(\nu) = \emptyset$, but for some $k \in J_{c^1}$, $B_k(\mu) \neq \emptyset$. Furthermore, $\mu(J_{c^1}) = \{c^1\} = \nu(J_{c^1})$. But then by *separability* of f, we obtain $\mu|_{N\setminus J_{c^1}} \succnsim_f^{E^1} \nu|_{N\setminus J_{c^1}}$ where $E^1 = E^{(q)}|_{(N\setminus J_{c_1}, C\setminus\{c^1\})}$. Note that in both cases we obtain $\mu|_{N\setminus J_{c^1}} \succsim_f^{E^1}$ where $E^1 = E^{(q)}|_{(N\setminus J_{c_1}, C\setminus\{c^1\})}$.

Then, in a similar fashion we continue with J_{c^2} and obtain $\mu|_{N\setminus(J_{c^1}\cup J_{c^2})} \succsim_f^{E^2} \nu|_{N\setminus(J_{c^1}\cup J_{c_2})}$ where $E^2 = E^{(q)}|_{(N\setminus(J_{c_1}\cup J_{c^2}), C\setminus\{c^1,c^2\})}$, and so on until we obtain $\mu|_{N\setminus(J_{c^1}\cup\cdots\cup J_{c^{q-1}})} \succsim_f^{E^{q-1}} \nu|_{N\setminus(J_{c^1}\cup\cdots\cup J_{c_{q-1}})}$ where $E^{q-1} = E^{(q)}|_{(N\setminus(J_{c_1}\cup\cdots\cup J_{c^{q-1}}), C\setminus\{c^1,\ldots,c^{q-1}\})}$.

Now, at the problem $E^{(q)}|_{(N\setminus(J_{c_1}\cup\cdots\cup J_{c^{q-1}}), C\setminus\{c^1,\ldots,c^{q-1}\})}$, for J_{c^q} we have $B_i(\nu|_{N\setminus(J_{c^1}\cup\cdots\cup J_{c_{q-1}})}) = \emptyset$ for all $j \in J_{c^q}$ and $B_{c^q}(\nu|_{N\setminus(J_{c^1}\cup\cdots\cup J_{c_{q-1}})}) = \emptyset$, but for some $k \in J_{c^q}$, $B_k(\mu|_{N\setminus(J_{c^1}\cup\cdots\cup J_{c_{q-1}})}) \neq \emptyset$. Furthermore, $\mu(J_{c^q}) = \{c^q\} = \nu(J_{c^q})$. But then, by *separability* of f we obtain

$$\mu|_{N\setminus(J_{c^1}\cup\cdots\cup J_{c^q})} \succnsim_f^{E^q} \nu|_{N\setminus(J_{c^1}\cup\cdots\cup J_{c_q})}$$

where $E^q = E^{(q)}|_{(N\setminus(J_{c_1}\cup\cdots\cup J_{c^q}), C\setminus\{c^1,\ldots,c^q\})}$. This is a contradiction since $I_{<k} \neq \emptyset$ (note that $k > 1$) and for all $j \in N\setminus(J_{c^1} \cup \cdots \cup J_{c^q})$, $\mu(j) = \nu(j)$. ■

Concluding the proof: Let $E = (N, C, R, q, \succeq) \in D$. We show by induction on k that, for any step k of the TTC algorithm at E, the assignment of an agent who is assigned an object at that step is the same as at $\varphi(E)$.

Base case: For each agent i who is assigned an object at Step 1 of the TTC algorithm at R, $\varphi_i(R) = TTC_i(R)$. This follows from Lemma 3.

Inductive step: Assume that for each agent i who is assigned an object at an earlier step than Step k of the TTC algorithm at R, $\varphi_i(R) = TTC_i(R)$. We will show that for each agent j who is assigned an object at Step k of the TTC algorithm at R, $\varphi_j(R) = TTC_j(R)$.

Let I_k denote the set of agents who are assigned objects at Step k of $TTC(R)$ and C_k denote the set of objects that are allocated at Step k of $TTC(R)$. Note that if for each $i \in I_k$, $\varphi_i(R) \in C_k$, then by *efficiency*, $\varphi_i(R) = TTC_i(R)$ for each $i \in I_k$.

Suppose that there exists $i_1 \in I_k$ such that $\varphi_{i_1}(R) \notin C_k$. Let $c_1 \in C_k$ be the object that is assigned at Step k of $TTC(R)$ and points to i_1 in Step k of $TTC(R)$. Let R'_{i_1} be a preference ordering for i_1 at which $TTC_{i_1}(R)$ is top-ranked and c_1 is second-ranked. By strategy-proofness, $\varphi_{i_1}(R'_{i_1}, R_{-i_1}) \neq TTC_{i_1}(R)$. By Lemma 4 and *strategy-proofness*, $\varphi_{i_1}(R'_{i_1}, R_{-i_1}) = c_1$.

Note that I_k is still the set of agents who are assigned objects at Step k of $TTC(R'_{i_1}, R_{-i_1})$ and C_k is still the set of objects that are allocated at Step k of $TTC(R'_{i_1}, R_{-i_1})$. Now, if for each $i \in I_k \setminus \{i_1\}$, $\varphi_i(R) \in C_k$, then *efficiency* would imply that $\varphi_i(R'_{i_1}, R_{-i_1}) = TTC_i(R'_{i_1}, R_{-i_1})$ for each $i \in I_k$, which would contradict $\varphi_{i_1}(R'_{i_1}, R_{-i_1}) \neq TTC_{i_1}(R) = TTC_{i_1}(R'_{i_1}, R_{-i_1})$. Thus, there exists $i_2 \in I_k \setminus \{i_1\}$ such that $\varphi_{i_2}(R) \notin C_k$. Let $c_2 \in C_k$ be the object that is assigned at Step k of $TTC(R)$ and points to i_2 in Step k of $TTC(R)$. Let R'_{i_2} be a preference ordering for i_2 at which $TTC_{i_2}(R)$ is top-ranked and c_2 is second-ranked. By strategy-proofness, $\varphi_{i_2}(R'_{i_1}, R'_{i_2}, R_{-\{i_1,i_2\}}) \neq TTC_{i_2}(R)$. By Lemma 4 and *strategy-proofness*, $\varphi_{i_2}(R'_{i_1}, R'_{i_2}, R_{-\{i_1,i_2\}}) = c_2$.

Continuing in a similar fashion, we identify a list of agents $(i_1, \ldots, i_m)$ and a preference profile $R' = (R'_{i_1}, \ldots, R'_{i_m}, R_{-\{i_1,\ldots,i_m\}})$ such that $\{i_1, \ldots, i_m\} \subseteq I_k$, $\varphi_i(R') \in C_k$ for each $i \in I_k$, and $\varphi_{i_m}(R') \neq TTC_{i_m}(R')$, which contradicts *efficiency* of φ.

Appendix B Independence of the properties

The examples below show that the three properties, *stability-preferred*, *separability*, and *consistency*, are independent for stability comparison methods.

Example 1 (Only stability-preferred violated) *Consider the following stability comparison f with $\gtrsim_f = \emptyset$, that is, for any problem E and any $\mu, \nu \in \mathcal{A}(E)$, μ and ν are incomparable in terms of $\gtrsim_f^E$, i.e. $\gtrsim_f^E = \emptyset$. Note that separability and consistency are vacuously satisfied, while stability-preferred is clearly violated.*

Example 2 (Only separability violated) *Consider the following stability comparison f. For any $E \in \mathcal{E}$ and $\mu, \nu \in \mathcal{A}(E)$, let $\mu \gtrsim_f^E \nu$ if and only if $B(\nu)$ is not a proper subset of $B(\mu)$, i.e., $B(\nu) \not\subsetneq B(\mu)$. Note that $\mu \not\gtrsim_f^E \nu$ if and only if $B(\mu) \subsetneq B(\nu)$.*

Clearly, stability-preferred is satisfied. To see that consistency is satisfied, take any unit-capacity problem E, and any μ and ν such that $\mu \gtrsim^E \nu$ and for some $\emptyset \neq N' \subseteq N$, $\mu(i) = \nu(i)$ for all $i \in N'$, $B_i(\mu) = B_i(\nu) \subseteq \mu(N')$ for all $i \in N'$, and $B_c(\mu) = B_c(\nu) \subseteq N'$ for all $c \in \mu(N') = \nu(N') = C'$. But then, $B(\nu|_{N\setminus N'}) \not\subsetneq B(\mu|_{N\setminus N'})$ and therefore $\mu|_{N\setminus N'} \gtrsim_f^{E|(N\setminus N', C\setminus C')} \nu|_{N\setminus N'}$.

To see that separability is violated, consider the following unit-capacity problem E where $N = \{1,2,3,4,5\}$ and $C = \{c_1, c_2, c_3, c_4, c_5\}$. Only the relevant top parts of the preference and priority profiles are depicted.

R_1	R_2	R_3	R_4	R_5	$\succeq_{c_1}$	$\succeq_{c_2}$	$\succeq_{c_3}$	$\succeq_{c_4}$	$\succeq_{c_5}$
c_2	c_3	c_3	c_4	c_1	1	3	2	3	
c_1	c_2	c_4		c_5	5	1	3	4	
c_5		c_2				2			

Consider

$$\mu = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ c_5 & c_2 & c_3 & c_4 & c_1 \end{pmatrix}$$

$$\nu = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ c_1 & c_3 & c_2 & c_4 & c_5 \end{pmatrix}$$

where $B(\mu) = \{(1,c_2),(1,c_5),(2,c_3)\}$ and $B(\nu) = \{(3,c_4)\}$. Let $N' = \{1,5\}$. Note that $\mu \gtrsim_f^E \nu$, $\mu(N') = \nu(N') = \{c_1, c_5\} = C'$, no agent in N' or no object in C' is involved in a blocking pair at ν, and $(1,c_1) \in B(\mu)$. However, $\nu|_{N\setminus N'} \gtrsim_f^{E(N\setminus N', C\setminus C')} \mu|_{N\setminus N'}$, implying that separability is violated.

Example 3 (Only consistency violated) *Consider the following stability comparison f. For any problem $E \in \mathcal{E}$ and $\mu, \nu \in \mathcal{A}(E)$, let $\mu \gtrsim_f^E \nu$ if and only if $B(\mu) = \emptyset$ or ($|B(\nu)| \geq 2$ and $|B(\mu)| \leq |B(\nu)|$). Note that $\mu \gnsim_f^E \nu$ if and only if $B(\mu) = \emptyset \neq B(\nu)$ or ($|B(\nu)| \geq 2$ and $B(\mu) < |B(\nu)|$). Also note that when $|B(\mu)| = |B(\nu)| = 1$, μ and ν are incomparable in terms of $\gtrsim^E$.*

By definition, stability-preferred is satisfied. To see that separability is satisfied, take any μ and ν such that $\mu \gtrsim_f^E \nu$, and take any $\emptyset \neq N' \subseteq N$ such that no agent in N' or no object in C' is involved in a blocking pair at ν, $\mu(N') = \nu(N') = C'$ and $B_i(\mu) \neq \emptyset$ for some $i \in N'$. Note that $|B(\nu|_{N\setminus N'})| \geq 2$ and $|B(\mu|_{N\setminus N'})| < |B(\nu|_{N\setminus N'})|$. Hence, $\mu|_{N\setminus N'} \gnsim_f^{E_{N\setminus N'}} \nu|_{N\setminus N'}$.

To see that consistency is violated, consider the following unit-capacity problem E where $N = \{1,2,3,4,5\}$ and $C = \{c_1, c_2, c_3, c_4, c_5\}$. Only the relevant top parts of the preference and priority profiles are depicted.

R_1	R_2	R_3	R_4	R_5	$\succeq_{c_1}$	$\succeq_{c_2}$	$\succeq_{c_3}$	$\succeq_{c_4}$	$\succeq_{c_5}$
c_2	c_2	c_4	c_4	c_5	1	1	3	3	4
c_1		c_3	c_5			2		4	5
			c_3						

Consider

$$\mu = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ c_1 & c_2 & c_3 & c_4 & c_5 \end{pmatrix}$$

$$\nu = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ c_1 & c_2 & c_4 & c_3 & c_5 \end{pmatrix}$$

where $B(\mu) = \{(1, c_2), (3, c_1)\}$ *and* $B(\nu) = \{(1, c_2), (4, c_5)\}$. *Let* $N' = \{1, 2\}$.

Note that $\nu \succsim_f^E \mu$, $\nu(i) = \mu(i)$ *for all* $i \in N'$, $B_i(\mu) = B_i(\nu) \subseteq \mu(N')$ *for all* $i \in N'$, *and* $B_c(\mu) = B_c(\nu) \subseteq N'$ *for all* $c \in \mu(N') = \nu(N') = C'$. *Yet,* $\nu|_{N \setminus N'}$ *and* $\mu|_{N \setminus N'}$ *are incomparable, implying that consistency is violated.*