UTS UNIVERSITY OF TECHNOLOGY SYDNEY

UTS Business School –
Department of Economics
Dr Chau Chak Wing Building
14 – 28 Ultimo Road, Ultimo NSW 2007

T: +61 2 9514 7713
economics@uts.edu.au

Department of Economics
Working Paper Series

'Stability against Robust Deviations in the Roommate Problem'

Daisuke Hirata [1]
Yusuke Kasuya [2]
Kentaro Tomoeda [3]

[1] Hitotsubashi University
[2] Kobe University
[3] University of Technology Sydney

No. 07 / 2019 **March 2020 (updated)**

Stability against Robust Deviations in the Roommate Problem

Daisuke Hirata* Hitotsubashi University

Yusuke Kasuya† Kobe University

Kentaro Tomoeda‡ University of Technology Sydney

This Version: March 11, 2020

Abstract

We propose a new solution concept in the roommate problem, based on the "robustness" of deviations (i.e., blocking coalitions). We call a deviation from a matching *robust up to depth k*, if none of the deviators gets worse off than at the original matching after any sequence of at most k subsequent deviations. We say that a matching is *stable against robust deviations (for short, SaRD) up to depth k*, if there is no robust deviation up to depth k. As a smaller k imposes a stronger requirement for a matching to be SaRD, we investigate the existence of a matching that is SaRD with a minimal depth k. We constructively demonstrate that a SaRD matching always exists for $k = 3$ and establish sufficient conditions for $k = 1$ and 2.

*d.hirata@r.hit-u.ac.jp

†kasuya@econ.kobe-u.ac.jp

‡Kentaro.Tomoeda@uts.edu.au

Contents

1 Introduction **1**
1.1 Related Literature 6

2 Preliminaries **7**
2.1 Party Permutation and Stable Partition 9

3 Main Results **11**
3.1 Proof Ideas 12

4 Discussions **20**
4.1 Relation to Other Solution Concepts 20
4.2 Further Discussions of our Concepts and Definitions 23

References **26**

A Construction of a SaRD Matching up to Depth $\frac{\#(N,\succ)-1}{2}$ **28**
A.1 Sufficient Conditions 28
A.2 Description of Algorithm A 32
A.3 Properties of Algorithm A 33

B Construction of a SaRD Matching up to Depth 3 **34**
B.1 Sufficient Conditions 34
B.2 Description of Algorithm B 37
B.3 Properties of Algorithm B 44

C Other Proofs **46**
C.1 Proof of Proposition 1 46
C.2 Proof of Proposition 2 46
C.3 Proof of Proposition 3 47

D Tightness of Theorems 1–2 with respect to σ **47**

E SaRD and Pareto Efficiency **51**

F Weak Stability against Robust Deviations **54**

G Histroty-Dependant Rational-Expectation Farsighted Stable Sets **57**

1 Introduction

Stability has been a central concept in various matching models. For instance, it has been one of the major desiderata for the design of two-sided matching markets, such as medical intern matching (Roth and Peranson, 1999) and school choice (Abdulkadiroglu and Sonmez, 2003). Further, it is also a primary solution concept in one-sided matching models, such as hedonic coalition formation (Bogomolnaia and Jackson, 2002) and network formation (Jackson, 2008).[1]

The *roommate problem*, which is a problem to partition finite agents into pairs (roommates) and singletons, is a simplest class of one-sided matching models. Despite its simplicity, however, the roommate problem may not possess a stable matching (Gale and Shapley, 1962); i.e., it is possible that at any (individually rational) matching outcome, there is a pair of agents who prefer each other to their current partners.[2] The non-existence of a stable outcome is not unique to the roommate problem but is a general obstacle in studying one-sided problems. Specifically, both coalition formation and network formation include the roommate problem as a special case, where respectively the size of each coalition is capped by two and each node can span at most one link. Therefore, these problems inherit the non-existence. Given its simplicity and connection to a broader class of problems, studying the roommate problem would be a natural first step to understand stability and related concepts in one-sided matching problems.[3]

In this paper, we propose a (class of) new stability concept(s) in the roommate problem and examine its existence property. The main idea is to find matchings that are free from "robust" deviations (i.e., blocking coalitions). We call a deviation from

[1] By "one-sided" models, we refer to those where any agent can be matched with any other. This is in contrast to "two-sided" models, where agents are partitioned into two sides and any match is between the two sides.

[2] Moreover, the proportion of problem instances (i.e., preference profiles) with no stable matching increases steeply as the number of agents increases (Gusfield and Irving, 1989; Pittel and Irving, 1994).

[3] Indeed, Klaus et al. (2010, p. 2219) write "roommate markets can be considered as an important benchmark for the development of solution concepts for matching, network and coalition formation models."

a matching robust up to depth k, if none of the deviators gets worse off than at the original matching after any sequence of at most k subsequent deviations. Then, we consider concepts that are weaker than the standard stability: a matching is *stable against robust deviations (for short, SaRD) up to depth k*, if there is no robust deviation up to depth k. Our main theorem shows that for any roommate problem, a SaRD matching up to depth $k \geq 3$ always exists. We also establish sufficient conditions for the existence of a SaRD matching up to depth $k = 1$ and 2.

To define our solution concept, we first differentiate potential deviations from a matching based on their "robustness." We say that a subset D of agents forms a deviation from an original matching μ if all agents in D can be strictly better off by rematching with each other. Suppose that a deviation D from μ leads to a new matching ν. If ν is not stable, which must be the case when no stable matching exists at all, the "original" deviation to ν may be followed by a second deviation, the second by a third, and so on. Figure 1 illustrates a "tree" of such deviation chains: ν has three possible deviations that lead to ν_1^1, ν_1^2, and ν_1^3, these in turn have further deviations to $\nu_2^1, \nu_2^2, \ldots, \nu_2^6$, and so on. Taking the possibility of subsequent deviations into account, we define the robustness of an original deviation as follows: a deviation is *robust up to depth k* if none of the deviators gets worse off than at the original matching after *any* sequence of $\kappa \leq k$ subsequent deviations. In the case of Figure 1, for instance, the deviation from μ to ν is robust up to depth 2 if none of the deviators gets worse off at *any* of the matchings $v_1^1, \ldots, v_1^3$ and $\nu_2^1, \ldots, \nu_2^6$ than at μ. It is robust up to depth 1 but not up to depth 2 if none in D gets worse off at any of $v_1^1, \ldots, v_1^3$ but at least one does at some of $\nu_2^1, \ldots, \nu_2^6$. When a deviation is robust up to depth k, the deviators are guaranteed to be better off unless sufficiently many (i.e., more than k) subsequent deviations follow.

A possible way to interpret our robustness concept is to suppose that agents have max-min preferences and search for the worst-case consequence of their deviation within those after k or less subsequent deviations. In such a scenario, potential deviators would agree to form a deviation if (and only if) it is robust up depth k. With

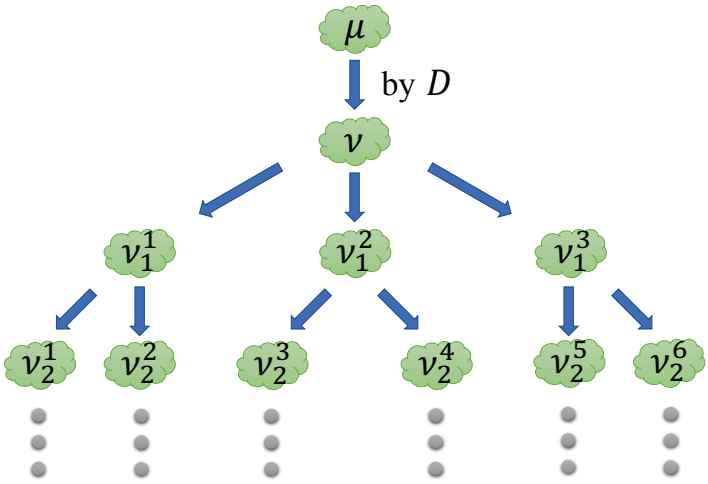


Figure 1: Tree of deviations

---

this interpretation, the depth k can be seen as the depth of reasoning, and the more sophisticated the agents are, the harder it is for them to agree on a possible deviation. Then it would be reasonable to argue that a deviation would be likely to realize when it is robust up to a large depth k, as it would be reachable even among extremely risk-averse and highly sophisticated agents.

For another interpretation that would be more broadly applicable, suppose next that forming a deviation takes a certain period of time and hence, at most one deviation can occur per period. This would be a reasonable presumption, for instance, in the context of business alliances, as it should take time to reach an agreement with a new partner and/or to dissolve an old partnership. With such a dynamic interpretation, the robustness of a deviation up to depth k means that the gain from it is guaranteed to last for at least k periods of time, no matter what happens in the future. To form a non-robust deviation, in contrast, the deviators must accept the risk of potential losses within a shorter time window. It is then natural to argue that potential deviators would not hesitate to form robust deviations while they might for non-robust ones. For a matching to remain long, therefore, a robust deviation up to k would be a more serious threat than non-robust deviations and those robust up to smaller k's.

Based on the idea that robust deviations are more likely to be realized than non-robust ones, we search for a matching that is free from the most serious deviations when any matching is subject to some deviations (i.e., when no stable matching exists). More specifically, we define a matching to be *stable against robust deviations* (henceforth, *SaRD*) *up to depth k* when no deviation from it is robust up to depth k. By definition, if a matching is SaRD up to depth k so is it up to any higher depth $k' > k$. Our objective is thus to investigate the existence of a matching that is SaRD up to as small depth k as possible.

To see how our concepts work in simple cases, suppose first that three agents a_1, a_2, and a_3 have a preference such that, respectively, $a_2 \succ_{a_1} a_3$, $a_3 \succ_{a_2} a_1$, and $a_1 \succ_{a_3} a_2$. If the initial matching is such that every agent is single, none can get strictly worse off after any sequence of voluntary deviations. That is, any deviation (e.g., the one by $D = \{a_1, a_2\}$) is robust up to any depth k and therefore, this initial matching is not SaRD up to any depth k. Now suppose instead that a_1 and a_2 are matched while a_3 is single at the initial matching. Then, a deviation is possible only by $D = \{a_2, a_3\}$, and after that, there is a unique subsequent deviation by $D' = \{a_1, a_3\}$. Notice that $a_2 \in D$ becomes single after D' deviates and hence strictly worse off than at the initial matching. That is, the original deviation by D is not robust up to depth 1 and the initial matching is SaRD up to depth 1.

Next suppose that there are five agents, from a_1 to a_5, and each a_i has a preference such that $a_{i+1} \succ_{a_i} a_{i-1}$ and all the other agents are unacceptable, where the subscripts are in modulo 5. As in the previous paragraph, a matching is not SaRD up to any depth k if it matches less than two pairs of agents. Suppose thus that there are two pairs, $\{a_1, a_2\}$ and $\{a_3, a_4\}$, and a_5 is left unmatched. Starting from this matching, the only possible deviation is by $D = \{a_4, a_5\}$, and thereafter, the unique subsequent deviations are first by $D_1 = \{a_2, a_3\}$ and then by $D_2 = \{a_1, a_5\}$. Notice that a_4 and a_5 remain matched to each other when D_1 deviates, while $a_4 \in D$ becomes single after D_2 follows. That is, the original deviation by D is robust up to depth 1 but not up to

depth 2; consequently, the initial matching is SaRD up to depth 2 but not up to depth 1. Similarly, if there are seven agents with a cyclic preference profile as above, a matching is SaRD up to depth 3 if it matches three pairs of "adjacent" agents (e.g., $\{a_1, a_2\}$, $\{a_3, a_4\}$, $\{a_5, a_6\}$), and no matching is SaRD up to depth 2. From these observations, one might expect that the depth k for which we can guarantee the existence of a SaRD matching would be increasing in the size of the preference cycle.

In fact, our main result demonstrates that we can construct a matching that is SaRD up to depth $k = 3$ for *any* roommate problem; i.e., with any number of agents and any preference profile. To see the key idea underlying our construction, now suppose that there are nine agents, from a_1 to a_9, and each a_i's preference is such that $a_{i+1} \succ_{a_i} a_{i-1}$ and all the others are unacceptable, where the subscripts are in modulo 9. If we match four pairs of agents, say $\{a_1, a_2\}, \ldots, \{a_7, a_8\}$, while leaving a_9 as single, it is SaRD up to depth 4 but not up to depth 3 or smaller, for similar reasonings as in the previous paragraphs. However, if we instead match only three pairs, $\{a_1, a_2\}$, $\{a_4, a_5\}$, and $\{a_7, a_8\}$, this matching is SaRD up to depth 2: For instance, if $D = \{a_2, a_3\}$ deviates, a_2 gets worse off after two subsequent deviations, first by $D_1 = \{a_5, a_6\}$ and then by $D_2 = \{a_3, a_4\}$. The point here is that matching as many agents as possible may not be necessarily optimal to eliminate robust deviations.[4] Combining this idea with the general structure called *party permutation* (Tan, 1991), we demonstrate that we can bound the depth of the most robust deviations to $k = 3$ even for more complicated preferences.

Although no matching is SaRD up to depth $k = 2$ for some problems as we have mentioned above, our construction also establishes sufficient conditions for the existence of a SaRD matching up to depth $k = 1$ and 2. These conditions can be seen as an extension of Tan's (1991) condition for the existence of a stable matching, as all of them can be parameterized by a single common parameter. Unlike Tan's, our conditions are not necessary, but they are tight in a certain sense as we will argue in Section 3 and

[4]Note that in this example both of the above two matchings are Pareto optimal. We further discuss the relation between our SaRD and Pareto efficiency in Appendix E.

Appendix D.

The rest of the paper is organized as follows: Section 1.1 briefly overviews the related literature. Section 2 introduces our model and key definitions. Section 3 provides the main results and the proof ideas. Section 4 presents further discussions of our concepts, including their relation to other solutions concepts. Appendices A–C contain the proofs. Appendices D–F discuss, respectively, the tightness of our conditions for $k = 1$ and 2, Pareto efficiency of our solution concepts, and an alternative definition of our solution. Appendix G considers the history-dependent rational-expectation farsighted stable set of Dutta and Vartiainen (2020) in our setup, so as to compare it with our solution.

1.1 Related Literature

In the literature, a number of studies have defined stability concepts based on chains of deviations and their final outcomes, in a similar spirit with ours. Among others, the most closely related is Barberà and Gerber (2003). They study the hedonic coalition formation, which generalizes the roommate problem, and propose a solution concept called *durability*. We share the spirit with them in distinguishing what we call robust deviations, and actually, in the roommate problem their durability coincides with our SaRD up to a sufficiently large depth k. However, we further differentiate robust deviations across k's and look for a SaRD matching up to a minimal depth, whereas Barberà and Gerber (2003) treat all deviation chains of any length as equally serious. The set of SaRD matchings up to depth 3 is generally smaller than that of durable matchings and hence, our concept can be seen as a refinement of durability. Relatedly, Troyan et al. (2018) propose in the school choice problem a solution concept called *essential stability*, which also corresponds to our SaRD with a sufficiently large k. It should be noted, however, that a stable matching always exists in the school choice problem and their motivation differs from ours.

While we investigate a static model with dynamic arguments as a possible inter-

pretation and motivation, Kadam and Kotowski (2018) and Kotowski (2015) explicitly study a dynamic marriage market, where agents have their preferences over the histories (i.e., sequences) of matched partners. They also define stability concepts for their dynamic setting, but it should be noted that their concepts reduce to the standard stability in the static setting. Also in a dynamic marriage market, Kurino (2019) proposes *credible stability*, which reduces in the static setting to a weaker version of our SaRD up to depth $k = 1$. We formally define this weaker concept and establish its existence in Appendix F.

Unsolvable roommate problems have been long studied in economics and other related fields, and several more solution concepts have been proposed. These include maximum stable matchings (Tan, 1990), almost stable matchings (Abraham et al., 2006), P-stable matchings (Inarra et al., 2008), absorbing sets (Iñarra et al., 2013), and Q-stable matchings (Biró et al., 2016). Each of those solutions focuses on a part of the properties that a stable matching satisfies, and extends it to unsolvable problems. In addition, some studies apply other general concepts than stability to the roommate problem; e.g., stochastic stability (Klaus et al., 2010) and farsighted stable sets (Klaus et al., 2011). The relation between our SaRD and other solution concepts will be discussed in more detail in Section 4.1.

2 Preliminaries

A *roommate problem* $(N, \succ)$ consists of a finite set N of agents and a profile $\succ = (\succ_a)_{a \in N}$ of strict preference relations over N. Given agent a's strict preference $\succ_a$, we write $b \succeq_a c$ to denote $[b \succ_a c \text{ or } b = c]$. We say that an agent a is *acceptable to* another agent b if $a \succ_b b$. A matching is a bijection $\mu : N \to N$ satisfying $\mu^2(a) = a$ for all $a \in N$. In the examples below, we also identify a matching with the partition it induces; e.g., when we write $\mu = \{\{a, b\}, \{c\}\}$, it refers to the matching defined by $\mu(a) = b$, $\mu(b) = a$, and $\mu(c) = c$. Given a subset $D \subseteq N$ of agents and two match-

ings μ and ν, we write $\nu \succ_D \mu$ if $\nu(a) \succ_a \mu(a)$ holds for all $a \in D$, and similarly, $\nu \succeq_D \mu$ if $\nu(a) \succeq_a \mu(a)$ holds for all $a \in D$. A matching μ is called *individually rational* if $\mu \succeq_N \mathrm{id}$, where id denotes the identity mapping over N. A matching μ is said to *leave no mutually-acceptable pairs of singles* if

$$[a \succ_b b \text{ and } b \succ_a a] \Longrightarrow [\mu(a) \neq a \text{ or } \mu(b) \neq b],$$

holds for all $a, b \in N$. This can be seen as a mild efficiency property, as a mutually-acceptable pair of singles implies Pareto inefficiency. Let us call a matching *regular* if it is individually rational and leaves no mutually-acceptable pairs of singles.

A subset D of agents, associated with a matching ν, is said to form a *deviation from* μ if they prefer ν to μ and can enforce the change from μ to ν in the sense that their new partners are also in D. More precisely, we call (D, ν) a deviation from μ and write $\nu \rhd_D \mu$, if (i) $\nu \succ_D \mu$, (ii) $a \in D \Rightarrow \nu(a) \in D$, (iii) $[b \notin D \text{ and } \mu(b) \in D] \Rightarrow \nu(b) = b$, and (iv) $c, \mu(c) \notin D \Rightarrow \nu(c) = \mu(c)$.[5] Notice that when μ is individually rational and $|D| = 2$, the identity of D pins down the unique matching ν such that (D, ν) can be a deviation from μ. More specifically, for $(\{a, b\}, \nu)$ to be a deviation from an individually rational μ, ν needs to be such that $\nu(a) = b$, $\nu(b) = a$, $\nu(c) = c$ for all $c \in \{\mu(a), \mu(b)\} - \{a, b\}$, and $\nu(d) = \mu(d)$ for all $d \notin \{a, b, \mu(a), \mu(b)\}$. Although we will not fully specify the associated ν when $|D| = 2$, it should thus cause no confusion. A matching μ is *stable* if there is no deviation (D, ν) such that $\nu \rhd_D \mu$.

Now we introduce our key concepts. A deviation (D, ν) from μ is called *robust up to depth* $k \in \mathbb{N}$, if $\nu_\kappa \succeq_D \mu$ holds for any sequence of deviations $(D_1, \nu_1), \ldots, (D_\kappa, \nu_\kappa)$ with $\kappa \leq k$ such that

$$\nu_\kappa \rhd_{D_\kappa} \nu_{\kappa-1} \rhd_{D_{\kappa-1}} \cdots \rhd_{D_2} \nu_1 \rhd_{D_1} \nu. \qquad (*)$$

[5]Part (iii) of this definition implicitly assumes that the partners of the members of D at μ are left single after the deviation. In Section 4.2.2, we discuss an alternative definition of a deviation that allows for instantaneous rematch among the agents who are left behind by D.

When no deviation from it is robust up to depth k, a matching μ is said to be *stable against robust deviations* (henceforce, *SaRD*) up to depth k.[6] By definition, if a deviation is robust up to depth k, then so is it up to any $k' < k$. Consequently, if a matching is SaRD up to depth k, then so is it up to any depth $k' > k$. Also by definition, with any depth k, a SaRD matching must be individually rational and leave no mutually-acceptable pairs of singles.

Proposition 1. *For any $k \geq 1$, if a matching μ is SaRD up to depth k, then it is regular.*

Proof. See Appendix C.1. ■

We relegate some further discussions concerning our definition of SaRD to Section 4. In Section 4.1, we discuss the relation between SaRD and other related solution concepts, including the bargaining set. In Section 4.2, we comment on possible criticism against our definitions.[7]

2.1 Party Permutation and Stable Partition

In this subsection, we introduce the concepts of a *party permutation* and of a *stable partition* (Tan, 1991), which we will heavily rely on in our analysis. A *permutation* is a bijection from N to itself. A permutation σ divides N into a finite number of cycles and hence, induces a partition $\mathscr{P}(\sigma)$ of N. Namely, $\{a_1, \ldots, a_n\} \subseteq N$ is a member of $\mathscr{P}(\sigma)$ if $\sigma^m(a_1) = a_{m+1}$ for all $m = 1, 2, \ldots, n-1$ and $\sigma^n(a_1) = a_1$. Throughout the rest of the paper, given a permutation σ over N, we let π denote its inverse σ^{-1} and call a pair (a, b) of agents *adjacent* if $\sigma(a) = b$ or $\sigma(b) = a$. We will focus on the following special class of permutations, which requires each $P \in \mathscr{P}(\sigma)$ to form a preference cycle:

Definition 1. A permutation $\sigma : N \to N$ is called a *semi-party permutation* if for each $P \in \mathscr{P}(\sigma)$, one of the following holds:

[6] In what follows, we use the acronym "SaRD" both as an adjective ("S" for stable) and as a noun ("S" for stability).

[7] In Appendices E and F, respectively, we also discuss the relationship between SaRD and Pareto efficiency and a weaker concept of SaRD.

- $|P| = 1$;
- $|P| = 2$ and $\sigma(a) \succ_a a$ for each $a \in P$; or
- $|P| \geq 3$ and $\sigma(a) \succ_a \pi(a) \succ_a a$ for each $a \in P$. □

Given a semi-party permutation σ and hence its inverse π, an agent $a \in N$ is said to be *superior* for another agent $b \in N$ when $a \succ_b \pi(b)$. When a is not superior for b (i.e., when $\pi(b) \succeq_b a$), then a is said to be *inferior* for b.[8] With this terminology, we can define an even more special subclass of semi-party permutations as follows:

Definition 2. A semi-party permutation σ is called a *party permutation* if the following holds: for any $a, b \in N$, if a is superior for b, then b is inferior for a. □

When σ is a party permutation, $\mathscr{P}(\sigma)$ is called a *stable partition*, and each of its elements a *party*. Given a party permutation σ, for each $a \in N$, let $P(a)$ denote the party a belongs to; i.e., $a \in P(a) \in \mathscr{P}(\sigma)$. A party P in a stable partition $\mathscr{P}(\sigma)$ is called *odd* (resp. *even*) if its cardinality is odd (resp. even). When it is a singleton, we call a party *solitary*. Note that when $\{a\} \in \mathscr{P}(\sigma)$ is a solitary party, b is acceptable to a if and only if b is superior for a.

While the definition of a stable permutation might look complicated, Tan (1991) shows that at least one exists for any problem and that odd parties are uniquely identified across all party permutations even when multiple exist:

Theorem (Tan, 1991)**.** *For any roommate problem $(N, \succ)$, at least one party permutation exists. If σ and σ' are both party permutations, then for any $P \subseteq N$ with $|P|$ being odd, $P \in \mathscr{P}(\sigma) \iff P \in \mathscr{P}(\sigma')$.*

For a problem $(N, \succ)$ with a party permutation σ, define $\#(N, \succ) \in \mathbb{N}$ by

$$\#(N, \succ) := \max\Big[\{|P| : P \in \mathscr{P}(\sigma) \text{ and } |P| \text{ is odd } \} \cup \{0\}\Big],$$

[8]Here we slightly modify Tan's (1991) original definition: when $\{a, b\} \in \mathscr{P}(\sigma)$, a and b are inferior for each other according to our definition, whereas they are neither superior nor inferior for each other according to Tan's. As this does not alter the definition of party permutations at all, Tan's (1991) results continue to hold with our definition.

which is independent of the choice of σ thanks to the above theorem. Namely, $\#(N, \succ)$ denotes the maximal size of odd parties in $(N, \succ)$ if there exists any, and it is set to zero otherwise. Put differently, $\#(N, \succ)$ is the length of the longest preference cycle among those involving an odd number of agents, since each party is a preference cycle by definition. With this notation, the existence of a stable matching can be characterized as follows. We will explain the basic ideas behind this theorem, which is a basis of our analysis, in Section 3.1 after we present our main results.

Theorem (Tan, 1991)**.** *A stable matching exists in a roommate problem $(N, \succ)$ if and only if $\#(N, \succ) \leq 1$.*

3 Main Results

In this section, we present our main results and explain the core ideas behind them, while we relegate the full proofs to Appendices A and B. Our first two results are on the existence of a SaRD matching up to depth $k = 1, 2$:

Theorem 1. *For any roommate problem $(N, \succ)$ such that $\#(N, \succ) \leq 3$, there exists a matching that is SaRD up to depth (at most)* 1*.*

Proof. This is a corollary of Propositions 4–5 in Appendix A. ■

Theorem 2. *For any roommate problem $(N, \succ)$ such that $\#(N, \succ) \leq 5$, there exists a matching that is SaRD up to depth (at most)* 2*.*

Proof. This is a corollary of Propositions 4–5 in Appendix A. ■

For each odd $n > 3$ (resp. odd $n > 5$), we can easily construct a problem $(N, \succ)$ such that $\#(N, \succ) = n$ and no matching is SaRD up to depth 1 (resp. depth 2). In this sense, the sufficient conditions in Theorems 1–2 are tight among those which depend only on $\#(N, \succ)$. In Appendix D, we further show that those conditions are almost tight among those depending only on σ.

The above observation, along with Tan's theorem, might suggest that it becomes harder to guarantee the existence of a SaRD matching up to a fixed depth k as $\#(N, \succ)$ grows larger. In fact, perhaps surprisingly, this is not the case. We can establish a uniform bound for the robustness of possible deviations, which applies to *any* problem $(N, \succ)$, as follows:

Theorem 3. *For any roommate problem $(N, \succ)$, there exists a matching that is SaRD up to depth (at most)* 3.

Proof. This is a corollary of Propositions 6–7 in Appendix B. ■

3.1 Proof Ideas

Before turning to the key ideas behind Theorems 1–3, it would be helpful to see how we can construct a stable matching when $\#(N, \succ) \leq 1$. To do so, fix a party permutation σ and suppose $\#(N, \succ) \leq 1$, which means that every party is either even or solitary. Then we can construct a stable matching as follows: For each a in an even party, match a to an agent "adjacent" to her with respect to σ (i.e., $\mu(a) \in \{\pi(a), \sigma(a)\}$); for each b in a solitary party, leave b as singles (i.e., $\mu(b) = b$). The following example illustrates the construction in a simple case.

Example 1. Let $N = \{a_1, \ldots, a_4, b\}$ and suppose that a party permutation is given by

$$\sigma = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & b \\ a_2 & a_3 & a_4 & a_1 & b \end{pmatrix},$$

where the right-hand side denotes $\sigma(a_1) = a_2$, $\sigma(a_2) = a_3$, and so on. Note that $\mathscr{P}(\sigma) = \{\{a_1, \ldots, a_4\}, \{b\}\}$. Then, the above construction leads to either $\mu_1 = \{\{a_1, a_2\}, \{a_3, a_4\}, \{b\}\}$ or $\mu_2 = \{\{a_2, a_3\}, \{a_4, a_1\}, \{b\}\}$. □

Although multiple matchings may satisfy the above conditions as in the example, they are all stable. The point here is that $\mu(i) \succeq_i \pi(i)$ holds for all $i \in N$, or equiva-

lently,

$$I_\mu^\circ := \{i \in N : \pi(i) \succ_i \mu(i)\}, \tag{1}$$

is empty. (Remember that $\pi(i) = i$ when $\{i\}$ is a solitary party.) This implies that if $b \succ_a \mu(a)$ for some agent b, then b is superior for a. For a pair of agents to form a deviation from μ, thus, each must be superior for the other. However, such a pair cannot exist by the definition of a party permutation, and hence, μ is stable.

When $\#(N, \succ) > 1$, conversely, the main problem is that I_μ° cannot be empty for any matching μ, and this is essentially why a stable matching fails to exist. Yet we can endow I_μ° with certain properties so as to bound the robustness of possible deviations. To explain those properties, given a matching μ, let

$$A_\mu := \{a \in N : P(a) > 1 \text{ and } \mu(a) \in \{\pi(a), \sigma(a)\}\},$$

denote the set of those who are matched to their "adjacent" agents with respect to σ, and $R_\mu := N - A_\mu$ the set of those who are matched to a "remote" partner (or remain single). Note that by definition, I_μ° is a subset of R_μ. To impose restrictions on I_μ°, thus, we first need to consider (i) how to divide N into A_μ and R_μ. Actually, in what follows we always put all the even party members into A_μ, as we did in the construction of a stable matching above. Hence our first task can be restated as (i') how to divide the odd party members into A_μ and R_μ. Second, we also need to consider (ii) how to match the agents in R_μ so as to bound the robustness of deviations from μ.

3.1.1 Proof Ideas for Theorems 1 and 2

When $\#(N, \succ) \leq 5$ holds, the division of N into A_μ and R_μ can be simple. For Theorems 1 and 2, we can match as many "adjacent" pairs as possible in each odd party (as well as in the even parties). In a party P with $|P| = 3$, we match arbitrary two agents in P with each other. Similarly, for a party P with $|P| = 5$, we match arbitrary four agents

as two adjacent pairs. That is, we can focus on matchings such that R_μ collects exactly one member from each odd party. Such matchings can be formally summarized as follows:

Property 0. *For each party $P \in \mathscr{P}(\sigma)$, $|P \cap R_\mu| \leq 1$.*

The main problem for $\#(N, \succ) \leq 5$ is how to match the agents among R_μ in order to minimize the robustness of deviations. We have to match mutually acceptable pairs among R_μ since any SaRD matching is regular (Proposition 1). However, we cannot arbitrarily do so, since how to match those pairs affect the robustness of possible deviations from the resulting matching. To see the point, suppose $D = \{a, b\}$ deviates from an original matching matching μ. Note that if a is single at μ, she cannot be worse off than that after any sequence of deviations. To check the robustness of the deviation by $\{a, b\}$, thus, we pick a with $\mu(a) \neq a$ and construct a (shortest) chain of subsequent deviations involving b so that a is worse off in the end than at the original μ. It is easier to do so if a is inferior for b since if so, b chooses a broader set of agents over a than otherwise. The following property guarantees that after possible relabeling, we can always pick a who is not single at μ and is inferior for b.[9]

Property 1. *For any $a, b \in N$, if a is superior for b and $\mu(b) = b$, then $\mu(a) \succ_a b$.*

In Appendix A, we provide an algorithm (Algorithm A) to construct a matching that satisfies regularity and Properties 0–1, and we show that its outcome is SaRD up to depth 1 and 2, respectively, when $\#(N, \succ) \leq 3$ and $\#(N, \succ) \leq 5$. The following example highlights our construction in a simple case, as well as showing that the lack of Property 1 actually gives rise to a more robust deviation.

Example 2. Consider a problem with $N = \{c_1, c_2, c_3, c_4, c_5, d_1, d_2, d_3, e_1, e_2, e_3\}$ and let

$$\sigma = \begin{pmatrix} c_1 & c_2 & c_3 & c_4 & c_5 & d_1 & d_2 & d_3 & e_1 & e_2 & e_3 \\ c_2 & c_3 & c_4 & c_5 & c_1 & d_2 & d_3 & d_1 & e_2 & e_3 & e_1 \end{pmatrix},$$

[9] See Lemma 1 in Appendix A.1.2.

Note that $\mathscr{P}(\sigma) = \Big\{\{c_1, c_2, c_3, c_4, c_5\}, \{d_1, d_2, d_3\}, \{e_1, e_2, e_3\}\Big\}$. Let $\succ$ be the preference profile specified as follows: For each $a \notin \{c_5, d_3, e_3\}$, her preference $\succ_a$ is such that $\sigma(a) \succ_a \pi(a)$ and all the other agents are unacceptable. For the remaining three agents, their preferences are given by

$$\begin{aligned} c_5: &\quad c_1 \succ_{c_5} c_4 \succ_{c_5} d_3 \succ_{c_5} e_3, \\ d_3: &\quad c_5 \succ_{d_3} d_1 \succ_{d_3} d_2, \qquad\qquad \text{and} \\ e_3: &\quad c_5 \succ_{e_3} e_1 \succ_{e_3} e_2, \end{aligned}$$

where all the unlisted agents are unacceptable. Note that σ as defined above is the unique party permutation for $(N, \succ)$. To construct a SaRD matching, first divide N into $A_\mu = \{c_1, c_2, c_3, c_4, d_1, d_2, e_1, e_2\}$ and $R_\mu = \{c_5, d_3, e_3\}$ so that Property 0 is satisfied. Note that among agents in R_μ, c_5 is a superior agent for both d_3 and e_3. In order to satisfy Property 1, c_5 must be matched to the preferred agent between $\{d_3, e_3\}$, who is d_3. Therefore,

$$\mu_1 = \Big\{\{c_1, c_2\}, \{c_3, c_4\}, \{c_5, d_3\}, \{d_1, d_2\}, \{e_1, e_2\}, \{e_3\}\Big\},$$

meets both the properties, while

$$\mu_2 = \Big\{\{c_1, c_2\}, \{c_3, c_4\}, \{c_5, e_3\}, \{d_1, d_2\}, \{d_3\}\{e_1, e_2\}\Big\},$$

violates Property 1. One can easily verify that μ_1 is SaRD up to depth 2. In contrast, μ_2 is not SaRD up to depth 2, since the deviation by $\{c_5, d_3\}$ is robust up to depth 2. □

3.1.2 Proof Ideas for Theorem 3

In the previous case, our construction was relatively simple in that

- we put as many agents as possible into A_μ (Property 0), and
- we consider the matching among R_μ after we finish fully partitioning N into A_μ

and R_μ.

For Theorem 3, we need to give up both of these features and our construction becomes more complicated. First of all, when $\#(N, \succ)$ is large, Property 0 does not help to minimize the robustness of deviations. The following example recap and slightly generalizes the point we have briefly discussed in the introduction.

Example 3. Consider the following class of problems: $N = \{a_1, \ldots, a_n\}$ and each a_i's preference is such that $a_{i+1} \succ_{a_i} a_{i-1}$ and all the others are unacceptable, where the subscripts are in modulo 9. First suppose $n = 9$. The matching

$$\mu_{(9)} = \Big\{\{a_1, a_2\}, \{a_3, a_4\}, \{a_5, a_6\}, \{a_7, a_8\}, \{a_9\}\Big\},$$

meets Property 0 but is not SaRD up to depth 3, since the deviation by $D = \{a_8, a_9\}$ is robust up to depth 3. In contrast,

$$\mu'_{(9)} = \Big\{\{a_1, a_2\}, \{a_3\}, \{a_4, a_5\}, \{a_6\}, \{a_7, a_8\}, \{a_9\}\Big\},$$

fails Property 0 but is SaRD up to depth 2. Next suppose $n = 11$. Similarly as above,

$$\mu_{(11)} = \Big\{\{a_1, a_2\}, \{a_3, a_4\}, \{a_5, a_6\}, \{a_7, a_8\}, \{a_9, a_{10}\}, \{a_{11}\}\Big\},$$

is not SaRD up to depth 3, but

$$\mu'_{(11)} = \Big\{\{a_1, a_2\}, \{a_3\}, \{a_4, a_5\}, \{a_6\}, \{a_7, a_8\}, \{a_9, a_{10}\}, \{a_{11}\}\Big\},$$

is SaRD up to depth 3. □

We thus omit Property 0 from our construction for Theorem 3, but the following weakening of it is still useful:

Property 2. *For any $a \in I^\circ_\mu$, $|P(a)|$ is odd and $\mu(\pi(a))$ is inferior for $\pi(a)$.*

When μ satisfies Property 0, $a \in I^\circ_\mu$ implies that $\pi(a)$ is matched to $\pi^2(a)$ and hence, Property2 is also satisfied. The virtue of Property 2 is in allowing us to focus on deviations with a particular structure. Suppose that D deviates from μ and $a \in D \cap I^\circ_\mu$. Then, Property 2 guarantees that such a deviation is robust up to depth 1 only if $\pi(a)$ is also a deviator; otherwise, $\{a, \pi(a)\}$ can form a further deviation, which makes $\nu(a) \in D$ strictly worse off. (This argument takes for granted that $\nu(a)$ was not single at μ, but it is without loss under regularity and Property 1.) With this property, thus, we can restrict our attention to deviations such that $a \in D \cap I^\circ_\mu$ implies $\pi(a) \in D$.

The second difficulty in the case of general $\#(N, \succ)$ is that we cannot fix the partition of N into A_μ and R_μ without considering how to match the agents among R_μ. From Example 3, one might expect that our strategy is to have one agent in R_μ between two adjacent pairs and to minimize the number of "successive" adjacent pairs. Although this is a part of our key ideas, it is not sufficient to find matching that is SaRD up to depth 3. The following example presents the problem that arises when we arbitrarily fix A_μ and R_μ.

Example 4. Consider a problem with $N = \{f_1, \ldots, f_{11}, g_1, g_2, g_3, h_1, h_2, h_3\}$ and let

$$\sigma = \begin{pmatrix} f_1 & \cdots & f_{11} & g_1 & g_2 & g_3 & h_1 & h_2 & h_3 \\ f_2 & \cdots & f_1 & g_2 & g_3 & g_1 & h_2 & h_3 & h_1 \end{pmatrix},$$

Note that $\mathscr{P}(\sigma) = \Big\{\{f_1, \ldots, f_{11}\}, \{g_1, g_2, g_3\}, \{h_1, h_2, h_3\}\Big\}$. Let $\succ$ be the preference profile specified as follows: For each $a \notin \{f_6, f_{11}, g_3, h_3\}$, her preference $\succ_a$ is such that $\sigma(a) \succ_a \pi(a)$ and all the other agents are unacceptable. The preferences of the remaining four agents are given by

$$f_6 : \quad f_7 \succ_{f_6} f_5 \succ_{f_6} g_3, \qquad f_{11} : \quad h_3 \succ_{f_{11}} f_1 \succ_{f_{11}} f_{10},$$
$$g_3 : \quad f_6 \succ_{g_3} g_1 \succ_{g_3} g_2, \quad \text{and} \quad h_3 : \quad h_1 \succ_{h_3} h_2 \succ_{h_3} f_{11},$$

where all the unlisted agents are unacceptable. Note that $\{f_6, g_3\}$ and $\{f_{11}, h_3\}$ are the (only) two "remote" mutually-acceptable pairs and that σ is the party permutation in $(N, \succ)$.

Now, suppose that we divide N as follows:

$$A_\mu = \{f_1, f_2, f_4, f_5, f_7, f_8, f_9, f_{10}, g_1, g_2, h_1, h_2\}, \text{ and}$$
$$R_\mu = \{f_3, f_6, f_{11}, g_3, h_3\}.^{10}$$

Since $\{f_6, g_3\}$ and $\{f_{11}, h_3\}$ are the only two mutually-acceptable pairs among agents in R_μ, regularity requires that these two pairs be matched. Given the partiaion of N as above, therefore,

$$\mu_1 = \Big\{\{f_1, f_2\}, \{f_3\}, \{f_4, f_5\}, \{f_6, g_3\}, \{f_7, f_8\}, \{f_9, f_{10}\}, \{f_{11}, h_3\}, \{g_1, g_2\}, \{h_1, h_2\}\Big\},$$

is the only matching that is consistent with the partition and is regular. However, μ_1 is SaRD only up to depth 4: If $D = \{f_5, f_6\}$ deviates, the shortest chain of subsequent deviations that makes one of D strictly worse off is those by $D_1 = \{h_2, h_3\}$, $D_2 = \{f_{10}, f_{11}\}$, $D_3 = \{f_8, f_9\}$, and $D_4 = \{f_6, f_7\}$. □

The problem of μ_1 in Example 4 is that once we first fix A_μ and R_μ, we necessarily have a set of agents $\{f_6, f_7, f_8, f_9, f_{10}, f_{11}\}$ such that (i) f_6 is matched to her inferior agent, (ii) $\{f_7, f_8\}$ and $\{f_9, f_{10}\}$ are adjacent pairs, and (iii) f_{11} is matched to her superior agent. Note that if $h_3 = \mu(f_{11})$ were instead inferior for f_{11}, $f_3 = \nu(f_6)$ would get worse off after the subsequent deviations by $D_1 = \{f_{10}, f_{11}\}$, $D_2 = \{f_8, f_9\}$, and $D_3 = \{f_6, f_7\}$; i.e., μ_1 would be SaRD up to depth 3. That is, we could circumvent the problem if we can avoid (iii) whenever (i)–(ii) hold. For this purpose, we need to construct a matching that also meets the following property:

[10] Note that the subpartition of $\{f_1, \ldots, f_{11}\}$ is analogous to $\mu'_{(11)}$ in Example 3, which is SaRD up to depth 3 in the simple, one-party case.

Property 3. *Let $a \in I_\mu^\circ$ be such that $\mu(\sigma(a)) = \sigma^2(a)$ and $\mu\left(\sigma^3(a)\right) = \sigma^4(a)$. If $|P(a)| = 7$, then $\mu\left(\sigma^5(a)\right) = \sigma^6(a)$. If $|P(a)| > 7$, then $\sigma^5(a) \in I_\mu^\circ \not\ni \sigma^6(a)$.*

In Appendix B, we show that regularity and Properties 1–3, along with one additional auxiliary property, are sufficient for a matching to be SaRD up to depth 3, and then provide an algorithm (Algorithm B) to compute a matching that meets all of those sufficient conditions. The key trick of this algorithm is in that we sequentially match both adjacent and remote pairs step by step, and hence, the partition of N into A_μ and R_μ is not fully determined until we go over those steps. While the entire procedure is complicated and we relegate it to the appendix, here we roughly illustrate how it works in the previous example.

Example 4 (Continued). Suppose that $(N, \succ)$ is the same as in Example 4 above. Our algorithm begins with finding a remote pair such that one of them is superior for the other. Let us fix such a pair, say, $\{f_6, g_3\}$. Then our algorithm matches the remote pair, $\{f_6, g_3\}$, as well as the following adjacent pairs in the parties that f_6 and g_3 belong to: $\{f_2, f_3\}$, $\{f_4, f_5\}$, $\{f_7, f_8\}$, $\{f_{10}, f_{11}\}$, and $\{g_1, g_2\}$. Next, we search again, among the unmatched agents, a remote pair such that one is superior for the other, but in this case there is no such pair among $\{f_1, f_9, h_1, h_2, h_3\}$. This allows us to match an arbitrary pair, say $\{h_1, h_2\}$, among the "untouched" party of $\{h_1, h_2, h_3\}$. In general, we proceed to match remote pairs among the rest, but in this case no pair among $\{f_1, f_9, h_3\}$ is mutually acceptable. Therefore, the final outcome of our algorithm is

$$\mu_2 = \Big\{\{f_1\}, \{f_2, f_3\}, \{f_4, f_5\}, \{f_6, g_3\}, \{f_7, f_8\}, \{f_9\}, \{f_{10}, f_{11}\}, \{g_1, g_2\}, \{h_1, h_2\}, \{h_3\}\Big\}.$$

The most robust deviation from μ_2 is by $D = \{f_1, f_{11}\}$, but this is not robust up to depth 3 because f_{11} becomes strictly worse off than at μ_2 after the subsequent deviations by $D_1 = \{f_5, f_6\}$, $D_2 = \{f_3, f_4\}$, and $D_3 = \{f_1, f_2\}$. One can easily check that the other deviations are not robust up to depth 3, either, and that μ_2 is SaRD up to depth 3. □

4 Discussions

4.1 Relation to Other Solution Concepts

4.1.1 Bargaining Set

Particularly with depth $k = 1$, our definition of SaRD matchings might remind readers of the bargaining set in cooperative game theory. In our definition, a deviation is robust if there are no further deviations that make an original deviator worse off, and a matching is SaRD if there is no robust deviation. In cooperative games, an objection is justified if it has no counterobjection, and an imputation is in the bargaining set if it has no justified objection. By definitions, our SaRD is a weakening of stability, whereas the bargaining set is a superset of the core, which is equivalent to the set of stable matchings in matching models. Given those similarities, it would be natural to ask how the SaRD matchings relate to the bargaining set.

To closely compare the two concepts, let us formally define Zhou's (1994) bargaining set in our setup.[11] An *objection* against a matching μ is a deviation (D, ν) from μ. A *counterobjection* against an objection (D, ν) is a pair (D', ν') such that

- $D' - D, D - D', D \cap D'$ are all non-empty,
- for all $i \in D'$, $\nu'(i) \neq \mu(i)$ implies $\nu'(i) \in D'$, and
- $\nu'(a) \succeq_a \mu(a)$ for all $a \in D' - D$ and $\nu'(b) \succeq_b \nu(b)$ for all $b \in D \cap D'$.

The similarity between our SaRD matchings and the bargaining set lies in that both require the existence of some (D', ν') that precludes a deviation (D, ν) from (or, an objection against) μ.

The key distinction, however, exists in the reference points with which (D', ν') is compared. On the one hand, in our definition of SaRD matchings, (D', ν') is a deviation from ν and hence, all the agents in D' compare ν and ν'. On the other hand, in the definition of the bargaining set, the agents in $D' - D$ compare μ and ν'.[12] Conse-

[11] For a more standard definition and characterization of Zhou's bargaining set in matching problems, see Klijn and Massó (2003) and Atay et al. (2019). Our definition below is equivalent to theirs.

[12] While there exist a number of different definitions of a bargaining set (e.g., Aumann and Maschler,

quently, the (set of) SaRD matchings and bargaining set are logically independent as we formally state below:

Proposition 2. *For any $k \geq 1$, the set of matchings that are SaRD up to depth k neither always includes nor is always included in the (Zhou) bargaining set.*

Proof. See Appendix C.2. ■

4.1.2 Farsightedly Stable Set

Our concept of SaRD might also remind readers of the farsighted stable set à la Harsanyi (1974), as condition $(*)$ in the definition of robust deviations on page 8 might appear to resemble indirect dominance in the definition of stable sets.[13] In relation to the farsighted stable set, we make three remarks here: First, the stable set is a set solution whereas ours is a pointwise (i.e., matching-wise) concept. Moreover, Klaus et al. (2011) establish in the roommate problem that a singleton is a farsighted stable set if and only if its unique element is a stable matching.[14] Therefore, although focusing on singletons can be a possible way to compare a set solution with a point solution, such an approach is not helpful to overcome the general non-existence of a stable matching in our setup.

Second, it should be noted that we can obtain exactly the same set of results even if we introduce "farsightedness" into our definitions. Specifically, let's say that a deviation (D, ν) is farsightedly-robust up to depth k, if $\nu_\kappa \succeq_D \mu$ for any sequence of deviations $(D_1, \nu_1), \ldots, (D_\kappa, \nu_\kappa)$ with $\kappa \leq k$ that satisfies $\nu_\kappa \succeq_{D_\lambda} \nu_{\lambda-1}$ for all $\lambda \in \{1, \ldots, \kappa\}$ (with $\nu_0 := \nu$) in addition to the original requirement $(*)$. Such a definition could be seen "farsighted" as the agents in D_λ also compare the final outcome (i.e., ν_κ) with the situation before they deviate (i.e., $\nu_{\lambda-1}$), while they myopically compare $\nu_{\lambda-1}$ and

1964; Mas-Colell, 1989) all of those we are aware of commonly require that a counterobjecction to be an objection against the original allocation (i.e., contain some comparison between ν' and μ). Hence our point here should apply to the general concept of bargaining sets, not only to the one by Zhou (1994).

[13] For the formal definitions of farsighted stable sets, see also Chwe (1994) and Ray and Vohra (2015).

[14] See also Ehlars (2007) and Mauleon et al. (2011) for related results in the marriage problem.

ν_λ in our original definitions. Actually, however, those alternative definitions do not affect our results and proofs at all. This is because whenever we consider a sequence of deviations, no agent deviates more than once along the sequence; that is, when we conclude that an original deviation is not robust up to depth k, it is also shown to be not farsightedly-robust up to depth k in the above sense.

Lastly, several recent studies (Ray and Vohra, 2019; Dutta and Vohra, 2017; Dutta and Vartiainen, 2020) propose new concepts of farsighted stable sets that incorporate dynamic consistency à la subgame perfection. Among them, the one by Dutta and Vartiainen (2020), history-dependent rational-expectation farsighted stable set (HREFS), is particularly relevant to the roommate problem, as it always exists in any finite game. In Appendix G, however, we provide a class of examples where the set of all individually rational matchings forms an HREFS. At least without further refinements, thus, the HREFS may be too inclusive and not necessarily useful in the context of the roommate problem.

4.1.3 P-stable matching

Inarra et al. (2008) propose the following concept of $\mathscr{P}$-stable matching, which is closely related to absorbing sets and stochastic stability in the roommate problem (Iñarra et al., 2013; Klaus et al., 2011):

Definition 3. Given a stable partition $\mathscr{P} = \mathscr{P}(\sigma)$, a matching μ is said to be $\mathscr{P}$-stable if μ satisfies Property 0 and $\mu(b) = b$ holds for all $b \in R_\mu$. □

In that they match as many "adjacent" pairs as possible (Property 0), $\mathscr{P}$-stable matchings are closely related to the outcomes of our Algorithm A. However, they differ in how to match the agents in R_μ: a $\mathscr{P}$-stable matching leaves all agents in R_μ unmatched although some of them may be mutually acceptable, whereas our algorithms necessarily remove such mutually-acceptable pairs of singles. As a result, $\mathscr{P}$-stable matchings match less agents than the outcomes of Algorithm A, which are SaRD up to depth $k = \frac{\#(N,\succ)-1}{2}$.

Proposition 3. *Suppose that* $\#(N, \succ) = 2k + 1$ *for some* $k \in \mathbb{N}$. *Then, for any* $\mathscr{P}$*-stable matching* μ'*, there exists a matching* μ *that is SaRD up to depth* k *and "includes"* μ' *in the sense that* $\mu'(a) = b \neq a$ *implies* $\mu(a) = b$ *for all* $a, b \in N$.

Proof. See Appendix C.3. ■

In contrast to $\mathscr{P}$-stable matchings and Algorithm A, our Algorithm B sometimes matches less adjacent pairs among odd parties to find a matching that is SaRD up to depth 3. Therefore, the matchings we construct for Theorem 3 do not "include" any $\mathscr{P}$-stable matching in general.

4.2 Further Discussions of our Concepts and Definitions

4.2.1 Consistency of the definition of SaRD matchings

One might argue that our concept of SaRD is inconsistent in that we try to exclude robust deviations while we allow non-robust subsequent deviations in defining robust deviations per se. In response to such a concern, we make two remarks. First, requiring consistency could lead to some subtlety, making it difficult for the solution to be a matching-wise concept. A natural way to require consistency would be to call a deviation "consistently robust" if the original deviators will be never worse-off after any subsequent deviations as long as those subsequent deviations are also "consistently robust." However, such a recursive definition might have multiple fixed points, each corresponding to *a different set of all "consistently robust" deviations*, and consequently, we would be unable to determine pointwise if a matching is "consistently SaRD" or not. Although we could jointly identify *multiple sets of all "consistently SaRD" matchings*, it would require something outside our model, such as beliefs of the agents, to choose a "right" one.

Second but not less importantly, we do not claim that a SaRD matching is fully immune to deviations or, in other words, that non-robust deviations would never realize. Instead we would argue, as did in the introduction, that robust deviations are more

likely to realize than the others and hence, that SaRD matchings are "less unstable" than the others. And our argument could still apply even if we define "consistently robust" deviations as above: The benefit from such a deviation is guaranteed under the hypothesis that only "consistently robust" deviations can follow. This hypothesis might be true if every agent is sophisticated enough to tell a deviation is "consistently robust" or not based on a shared criterion. However, even if an agent herself is sophisticated, she could be unsure if the others are also sophisticated. Further, even if she believes the others to be sophisticated as well, she could be still unsure what criteria of "consistent robustness" they adopt, since there could be multiple of them as argued above. For an agent facing such ambiguities, a deviation would be less secure when it is "consistently robust" than when it is robust in our sense. Our strategy in this study is to eliminate deviations that would be the most secure and likely to realize.

4.2.2 Definition of Deviations

Our definition in Section 2 requires a deviation (D, ν) from μ to satisfy $\nu(i) = i$ if $\mu(i) \in D$. That is, we implicitly assume that the agents who are left behind by D remain single at ν, while one might argue those agents could instantaneously rematch among themselves. To be concrete, let us call (D, ξ) a *deviation with (instantaneous) rematch* from μ and write $\xi \rhd_D^* \mu$ if (i) $\nu \succ_D \mu$, (ii) $a \in D \Rightarrow \nu(a) \in D$, (iii') $[b \notin D \text{ and } \mu(b) \in D] \Rightarrow \mu(\nu(b)) \in D$, and (iv) $c, \mu(c) \notin D \Rightarrow \nu(c) = \mu(c)$. We can also define the robustness of a deviation (with rematch) and the SaRD property using $\rhd^*$ instead of $\rhd$.

We make two remarks on such alternative definitions. First, once we fix an initial deviation (with or without rematch) from an original matching, its robustness measured by depth k is independent of whether we allow rematch or not in subsequent deviations, i.e., whether we use $\rhd$ or $\rhd^*$. This is because, if an original deviator is worse off after a subsequent deviation and rematch, she must be (even) worse off before the rematch after the subsequent deviation. Therefore, the issue here is whether

or not to allow instantaneous rematch for initial deviations. Second, allowing rematch for initial deviations weakly decreases the stability of a matching measured by depth k, since a deviation with rematch can be robust up to a larger depth than any deviations without rematch. This is because the instantaneous rematch (weakly) shrinks the set of possible subsequent deviations. To be more concrete, suppose that $\nu \rhd_D \mu$, $\xi \rhd^*_D \mu$, and $\nu(i) = \xi(i)$ for all $i \in D$. That is, ν and ξ differ only in that the agents in $\mu(D)$ are left single at ν while they are rematched among themselves at ξ. Since $\xi(j) \succeq_j \nu(j)$ for $j \in \mu(D)$, there may exist ξ_1 such that $\xi_1 \rhd^*_{D_1} \nu$ for some D_1 but not $\xi_1 \rhd^*_{D_1} \xi$ for any D_1. This is why (D, ξ) may be more robust than (D, ν). However, this merely means that a deviation may become more robust if the deviators can enforce a a particular way of rematch among those who they leave behind. Unless we presume such enforcement powers, thus, the alternative definitions based on $\rhd^*$ is against our spirit in this study, which is to measure the robustness of a deviation based on worst-case scenarios for the deviators.

Acknowledgments

We are particularly grateful to Benjamin Balzer for detailed comments. We also thank Isa Hafalir, Michihiro Kandori, Mamoru Kaneko, Bettina Klaus, Fuhito Kojima, Akihiko Matsui, Manabu Toda, Alvin E. Roth, Zaifu Yang, and seminar participants at GETA 2009, SWET 2009, Waseda University, Hitotsubashi University, the 14th Meeting of the Society for Social Choice and Welfare, the Lisbon meetings in Game Theory and Applications #10, Sydney EconCS Workshop 2019, and MATCH-UP 2019 for helpful discussions. Hirata gratefully acknowledges financial support from JSPS KAKENHI (#16K17081).

References

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

ABRAHAM, D. J., P. BIRÓ, AND F. MANLOVE DAVID (2006): ""Almost Stable" Matchings in the Roommates Problem," in *Approximation and Online Algorithms: Third International Workshop, WAOA 2005*, ed. by T. Erlebach and G. Persiano, Springer Berlin Heidelberg, 1–14.

ATAY, A., A. MAULEON, AND V. VANNETELBOSCH (2019): "A Bargaining Set for Roommate Problems," *mimeo*.

AUMANN, R. J. AND M. MASCHLER (1964): "The Bargaining Set for Cooperative Games," in *Advances in Game Theory*, ed. by M. Dresher, L. S. Shapley, and A. W. Tucker, Princeton University Press, Princeton, 443–476.

BARBERÀ, S. AND A. GERBER (2003): "On Coalition Formation: Durable Coalition Structures," *Mathematical Social Sciences*, 45, 185–203.

BIRÓ, P., E. IÑARRA, AND E. MOLIS (2016): "A new solution concept for the roommate problem: $\mathscr{Q}$-stable matchings," *Mathematical Social Sciences*, 79, 74–82.

BOGOMOLNAIA, A. AND M. O. JACKSON (2002): "The Stability of Hedonic Coalition Structures," *Games and Economic Behavior*, 38, 201–230.

CHWE, M. S.-Y. (1994): "Farsighted Coalitional Stability," *Journal of Economic Theory*, 63, 299–325.

DUTTA, B. AND H. VARTIAINEN (2020): "Coalition Formation and History Dependence," *Theoretical Economics*, 15, 159–197.

DUTTA, B. AND R. VOHRA (2017): "Rational Expectations and Farsighted Stability," *Theoretical Economics*, 12, 1191–1227.

EHLARS, L. (2007): "Von Neuman-Morgenstern Stable Sets in Matching Problems," *Journal of Economic Theory*, 134, 537–547.

GALE, D. AND L. S. SHAPLEY (1962): "College Admissions and the Stability of Marriage," *American Mathematical Monthly*, 69, 9–15.

GUSFIELD, D. AND R. W. IRVING (1989): *The Stable Marriage Problem: Structure and Algorithms*, MIT Press.

HARSANYI, J. C. (1974): "An Equilibrium-Point Interpretation of Stable Sets and a Proposed Alternative Definition," *Management Science*, 20, 1472–1495.

IÑARRA, E., C. LARREA, AND E. MOLIS (2013): "Absorbing sets in roommate problems," *Games and Economic Behavior*, 81, 165–178.

INARRA, E., C. LARREA, AND E. MOLIS (2008): "Random Paths to *P*-Stability in the Roommate Problem," *International Journal of Game Theory*, 36, 461–471.

JACKSON, M. O. (2008): *Social and Economic Networks*, Princeton University Press.

KADAM, S. V. AND M. H. KOTOWSKI (2018): "Multi-Period Maching," *International Economic Review 59*, 1927–1947.

KASUYA, Y. AND K. TOMOEDA (2012): "Credible Stability in the Roommate Problem," *mimeo*.

KLAUS, B., F. KLIJN, AND M. WALZL (2010): "Stochastic Stability for Roommate Markets," *Journal of Economic Theory*, 145, 2218–2240.

——— (2011): "Farsighted Stability for Roommate Markets," *Journal of Public Economic Theory 13*, 921–933.

KLIJN, F. AND J. MASSÓ (2003): "Weak Stability and a Bargaining Set for the Marriage Model," *Games and Economic Behavior*, 42, 91–100.

KOTOWSKI, M. H. (2015): "A Note on Stability in One-to-One, Multi-Period Matching Markets," *mimeo*.

KURINO, M. (2019): "Credibility, Efficiency, and Stability: A Theory of Dynamic Matching Markets," *Japanese Economic*, Review, forthcoming.

MAS-COLELL, A. (1989): "An Equivalence Theorem for a Bargaining Set," *Journal of Mathematical Economics*, 18, 129–139.

MAULEON, A., V. J. VANNETELBOSCH, AND W. VERGOTE (2011): "Von Neumann-Morgenstern Farsightedly Stable Sets in Two-Sided Matching," *Theoretical Economics*, 6, 499–521.

PITTEL, B. G. AND R. W. IRVING (1994): "An Upper Bound for the Solvability Probability of a Random Stable Roommates Instance," *Random Structures and Algorithms*, 5, 465–486.

RAY, D. AND R. VOHRA (2015): "The Farsighted Stable Set," *Econometrica*, 83, 977–1011.

——— (2019): "Maximality in the Farsighted Stable Set," *Econometrica 87*, 1763–1779.

ROTH, A. E. AND E. PERANSON (1999): "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design," *American Economic Review*, 89, 748–780.

TAN, J. J. M. (1990): "A Maximum Stable Matching for the Roommate Problem," *BIT*, 29, 631–640.

——— (1991): "A Necessary and Sufficient Condition for the Existence of a Complete Stable Matching," *Journal of Algorithms*, 12, 154–178.

TROYAN, P., D. DELACRÉTAZ, AND A. KLOOSTERMAN (2018): "Efficient and Essentially Stable Assignments," *mimeo*.

ZHOU, L. (1994): "A New Bargaining Set of an N-Person Game and Endogeneous Coalition Formation," *Games and Economic Behavior*, 6, 512–526.

A Construction of a SaRD Matching up to Depth $\frac{\#(N,\succ)-1}{2}$

In this appendix, we demonstrate how we can always construct, and thereby guarantee the existence of, a matching that is SaRD up to depth $\frac{\#(N,\succ)-1}{2}$. Note that Theorems 1–2 are a special case of the results in this appendix. In A.1, we first establish a set of sufficient conditions for a matching to be SaRD up to depth $\frac{\#(N,\succ)-1}{2}$. We provide in A.2 an algorithm to compute a matching for an arbitrarily given problem $(N,\succ)$, and we then show in A.3 that its outcome always satisfies the aforementioned sufficient conditions.

A.1 Sufficient Conditions

The goal of this subsection is to establish the following proposition:

Proposition 4. *Suppose* $\#(N,\succ) = 2k+1$ *for some* $k \in \mathbb{N}$ *and that* μ *is a regular matching satisfying Properties 0 and 1 (introduced in Section 3.1). Then,* μ *is SaRD up to depth* k*.*

To prove this proposition, we first introduce some more notation in A.1.1 and then establish a few implications of Properties 1–2 in A.1.2. Since Property 0 implies Property 2, the proof of Proposition 4 in A.1.3 can rely on those implications.

A.1.1 Preliminaries

Before we establish the implications of Properties 1–2, here we introduce some more notation to concisely state them: Taking a deviation (D,ν) from μ as given, let $S_\nu := \{a \in N : \nu(a) \succ_a \pi(a)\}$ be the set of agents who are matched to their superior agents

at ν. Note that if $\nu \rhd_D \mu$ and $a \in D \cap S_\nu$, then $\nu(a)$ must be a member of I°_μ, which has been defined in equation (1). Through this relation, the properties with respect to I°_μ have implications on $D \cap S_\nu$. Divide $D \cap S_\nu$ into two as follows:

$$Cy := \{a \in D \cap S_\nu : (\pi \circ \nu)^t(a) \in S_\nu \text{ for all } t \in \mathbb{N}\}, \text{ and}$$
$$Ch := (D \cap S_\nu) - Cy.$$

Note that by the finiteness of N,

$$[a \in Cy] \Longrightarrow [\text{there exists } t^* \in \mathbb{N} \text{ such that } (\pi \circ \nu)^{t^*}(a) = a],$$

where t^* becomes 1 when $\nu(a) = \sigma(a)$. That is, $a \in Cy$ means that $\pi \circ \nu$ forms a cycle within S_ν that involves a. In contrast, $a \in Ch$ implies $(\pi \circ \nu)^{t'}(a) \notin S_\nu$ for some t'; i.e., the chain induced by $\pi \circ \nu$ gets outside of S_ν before it forms a cycle.

A.1.2 Implications of Properties 1–2

The first lemma is a key implication of Property 1. It guarantees that for any deviation (D, ν), there exists some agent $a \in D \cap I^\circ_\mu$. Consequently, the other properties on μ regarding I°_μ become relevant.

Lemma 1. *Let μ be a regular matching satisfying Property 1, and suppose that $\nu \succ_E \mu$ where $E = \{a, b\}$ and $\nu(a) = b$. Then, at least one of the following holds: (i) $a \in I^\circ_\mu$, b is an inferior agent for a, and $\mu(b) \neq b$; and (ii) $b \in I^\circ_\mu$, a is an inferior agent for b, and $\mu(a) \neq a$.*

Proof. First, by the definition of a party permutation, either a is inferior for b or b is inferior for a (or both). Second, μ's regularity implies that at least one of $\mu(a) \neq a$ and $\mu(b) \neq b$ must hold.[15] Third, $\nu \succ_E \mu$ and Property 1 imply *both* [1] either a is inferior for b or $\mu(b) \neq b$, and [2] either b is inferior for a or $\mu(a) \neq a$. Combining those claims altogether, we can conclude that at least one of the following holds: [i] a is inferior for

[15]Note that μ's individually rationality and $\nu \succ_E \mu$ imply that a and b are mutually acceptable.

b and $\mu(a) \neq a$, and [ii] b is inferior for a and $\mu(b) \neq b$.

If $a \notin I^{\circ}_{\mu}$ and b is inferior for a, it follows that $\mu(a) \succeq_a \pi(a) \succeq_a b = \nu(a)$, but this is a contradiction to the assumption of $\nu \succ_E \mu$. Therefore, $a \in I^{\circ}_{\mu}$ if b is inferior for a, and symmetrically, $b \in I^{\circ}_{\mu}$ if a inferior for b. Combined with the conclusion of the previous paragraph, these complete the proof. ■

Next is a useful, albeit immediate, consequence of the previous lemma. It substantially simplifies our proof to bound the robustness of a deviation ν from μ. Specifically, suppose that $a \in D \cap S_\nu$ and $\nu_\kappa \rhd_{D_\kappa} \cdots \nu_1 \rhd_{D_1} \nu$, where $\nu(a) \in D_\kappa$ and $\nu_\kappa(a) = a$. Then, the following lemma guarantees that a prefers $\mu(a) \neq a$ to $\nu_\kappa(a) = a$, and thereby that ν is not robust up to depth κ.

Lemma 2. *Suppose $\nu \rhd_D \mu$, where μ is a regular matching satisfying Property 1. If $a \in S_\nu$, then $\mu(a) \neq a$.*

Proof. If $a \notin D$, the assumption of $a \in S_\nu$ means $\mu(a) = \nu(a) \succ_a \pi(a) \succeq_a a$, which implies $\mu(a) \neq a$. If $a \in D$ and hence $\nu \succ_{\{a,\nu(a)\}} \mu$, $\mu(a) \neq a$ follows from $a \in S_\nu$ and Lemma 1. ■

Next we turn to the implications of Property 2 on the structure of *Cy* and *Ch*.

Lemma 3. *Suppose $\nu \rhd_D \mu$, where μ is a regular matching satisfying Property 2. If $a \in D \cap S_\nu$ and $(\pi \circ \nu)(a) \in S_\nu$, then, $(\pi \circ \nu)(a) \in D$.*

Proof. Notice that $a \in D \cap S_\nu$ implies $\nu(a) \in D - S_\nu$ and hence $\nu(a) \in I^{\circ}_{\mu}$. By Property 2, $(\pi \circ \nu)(a)$ should be matched to an inferior agent at μ. Thus, $(\pi \circ \nu)(a) \in D$ is necessary for $(\pi \circ \nu)(a) \in S_\nu$ to hold. ■

Lemma 4. *Suppose $\nu \rhd_D \mu$, where μ is a regular matching satisfying Property 2. If Ch is nonempty, then there exists $a \in Ch$ such that $(\pi \circ \nu)(a) \notin S_\nu$.*

Proof. This is an immediate corollary of Lemma 3. ■

Lemma 5. *Suppose $\nu \rhd_D \mu$, where μ is a regular matching satisfying Properties 1 and 2 (with respect to a same party permutation σ). Then, ν is robust up to depth 1 only if $Ch = \varnothing \neq Cy$.*

Proof. Suppose $\nu \rhd_D \mu$, where μ is a regular matching satisfying Properties 1–2. We show that ν is not robust up to depth 1 if $Ch = \varnothing \neq Cy$ fails to hold, which we divide into two subcases. First, suppose that $D \cap S_\nu = \emptyset$. By the regularity of μ, for any $a, \nu(a) \in D$, either $\mu(a) \neq a$ or $\mu(\nu(a)) \neq \nu(a)$. Without any loss, suppose $\mu(\nu(a)) \neq \nu(a)$. On the one hand, a prefers $\pi(a)$ to $\nu(a)$ as $a \notin S_\nu$ by the assumption of $D \cap S_\nu = \emptyset$. On the other hand, $\pi(a)$ also prefers a to $\nu(\pi(a))$, as

- if $\pi(a) \in D$, $\pi(a) \notin S_\nu$ by the assumption of $D \cap S_\nu = \emptyset$, and
- otherwise, $\nu(\pi(a)) \in \{\pi(a), \mu(\pi(a))\}$ and $\mu(\pi(a))$ is inferior by Property 2.

Therefore, we can construct a further deviation ν' by matching a and $\pi(a)$, so that $\nu(a) \in D$ prefers $\mu(\nu(a)) \neq \nu(a)$ to $\nu'(\nu(a)) = \nu(a)$. That is, the original deviation ν is not robust up to depth 1.

Second, suppose that $Ch \neq \varnothing$. By Lemma 4, there exists $a \in Ch$ such that $(\pi \circ \nu)(a) \notin S_\nu$. Lemma 2 then implies $\mu(a) \neq a$ and thus, it suffices to establish a further deviation involving $\nu(a)$. As $a \in S_\nu$ and $(\pi \circ \nu)(a) \notin S_\nu$, $\nu(a)$ and $\pi(\nu(a))$ prefer each other to their partners at ν. We can thus construct a further deviation ν' from ν by matching $\nu(a)$ and $\pi(\nu(a))$ so that $a \in D$ prefers $\mu(a) \neq a$ to $\nu'(a) = a$. That is, the original deviation ν is not robust up to depth 1. ■

A.1.3 Proof of Proposition 4

Take an arbitrary deviation (D, ν) from μ. Suppose that $Ch = \varnothing \neq Cy$, as otherwise (D, ν) is not robust up to depth 1 by Lemma 5. Fix an arbitrary $b \in \nu(Cy) \subset I^\circ_\mu$. By Property 0, then, $\mu\left(\sigma^{2j}(b)\right) = \sigma^{2j-1}(b)$ holds for each $j = 1, \ldots, \frac{|P(b)|-1}{2}$.

To begin, let ℓ be the smallest positive integer such that $\nu\left(\sigma^{2\ell}(b)\right) \neq \sigma^{2\ell-1}(b)$. Such ℓ exists since $b \in \nu(Cy)$ implies $\pi(b) \in D$, while $\pi(b) \equiv \sigma^{|P(b)|-1}(b)$ is matched to $\sigma^{|P(b)|-2}(b)$ at μ; that is, $\ell = \frac{|P(b)|-1}{2}$ meets the condition. Further, $\sigma^{2\ell-1}(b)$ cannot be a member of D, and thus, she must be single at ν. To see this, suppose otherwise that

$\sigma^{2\ell-1}(b) \in D$. Then, since she is matched at μ to a superior partner, $\sigma^{2\ell}(b)$, so is she at ν; i.e., $\sigma^{2\ell-1}(b) \in D \cap S_\nu$. By the assumption of $Ch = \varnothing$, it follows that $\sigma^{2\ell-1}(b) \in Cy$ and thus, by the definition of Cy, that $\sigma^{2\ell}(b) \in \nu(Cy) \subseteq D - S_\nu$. This, however, is a contradiction, because $D - S_\nu \subseteq I^\circ_\mu$ whereas $\mu\left(\sigma^{2\ell}(b)\right) = \sigma^{2\ell-1}(b)$ implies $\sigma^{2\ell}(b) \notin I^\circ_\mu$.

Given $\sigma^{2\ell-1}(b)$ is single at ν and $\nu\left(\sigma^{2j}(b)\right) = \sigma^{2j-1}$ for all $j < \ell$, we can construct $\nu_1, \nu_2, \ldots, \nu_\ell$ by matching $D_1 = \{\sigma^{2\ell-1}(b), \sigma^{2\ell-2}(b)\}$, $D_2 = \{\sigma^{2\ell-3}(b), \sigma^{2\ell-4}(b)\}$, …, $D_\ell = \{\sigma(b), b\}$ so that $\nu_\ell \rhd_{D_\ell} \nu_{\ell-1} \rhd_{D_{\ell-1}} \ldots \rhd_{D_1} \nu$. Note that $\nu(b) \in D$ is single at D_ℓ but not at the original μ by Lemma 2. That is, $\nu(b)$ prefers μ to ν_ℓ, and the deviation (D, ν) is not robust up to depth ℓ. Since $\ell \leq \frac{|P(b)|-1}{2} \leq k$ by definition, the proof is complete. ■

A.2 Description of Algorithm A

Taking a problem $(N, \succ)$ and a party permutation σ as given, construct a matching μ as follows. To simplify the description, we write "define $\mu(a) := b$," when it should read as "define $\mu(a) := b$ and $\mu(b) := a$." The whole procedure is divided into two phases.

A.2.1 Phase 1 of Algorithm A

For each non-solitary party $P \in \mathscr{P}(\sigma)$, arbitrarily fix its member $a \in P$ and define $\mu\left(\sigma^{2j-1}(a)\right) = \sigma^{2j}(a)$ for each $j \in \{1, \ldots, \lfloor |P|/2 \rfloor\}$, where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x.

A.2.2 Phase 2 of Algorithm A

Let Λ_0 be the set of agents who are not matched in Phase 1, including the members of the solitary parties. Arbitrarily order the members of Λ_0 as $x_1, \ldots, x_T$ and iterate the following steps.

Step $t \leq T$: If $x_t \notin \Lambda_{t-1}$, then proceed to step $t+1$. Otherwise, let

$$\widetilde{\Sigma}_t := \{y \in \Lambda_{t-1} : x_t \text{ is superior for } y \text{ and } y \text{ is acceptable for } x_t\}.$$

If $\widetilde{\Sigma}_t$ is empty, let $\Lambda_t := \Lambda_{t-1}$ and proceed to step $t+1$ without defining $\mu(x_t)$. Otherwise, define $\mu(x_t) := y_t$, where y_t is the best partner for x_t among $\widetilde{\Sigma}_t$ (i.e., $y_t \succeq_{x_t} y$ for all $y \in \widetilde{\Sigma}_t$), and proceed to step $t+1$ with $\Lambda_t := \Lambda_{t-1} - \{x_t, y_t\}$.

Step $t > T$: If there exists a mutually-acceptable pair $(z, w) \in \Lambda_{t-1} \times \Lambda_{t-1}$, then let $\mu(z) := w$ and proceed to step $t+1$ with $\Lambda_t := \Lambda_{t-1} - \{z, w\}$. Otherwise, proceed to the final step with $\Lambda_F := \Lambda_{t-1}$.

Final Step: For any $a \in \Lambda_F$, define $\mu(a) := a$.

A.3 Properties of Algorithm A

Now we show that the outcomes of the above algorithm indeed satisfy the sufficient conditions of Proposition 4.

Proposition 5. *Any outcome μ of Algorithm A is a regular matching satisfying Properties 0 –1 (with respect to the party permutation σ fixed at the beginning of the algorithm).*

Proof. It is straightforward to check regularity and Property 0: First, μ is individually rational as the algorithm only matches mutually-acceptable pairs. Second, the second half of Phase 2 precludes any mutually-acceptable pair of singles. Third, the procedures of Phase 1 ensure Property 0.

To check Property 1, suppose a is superior for b and $\mu(b) = b$. Remember that b must be inferior for a, since no pair of agents are superior to each other, and that $\pi(a) \succeq_a b$ by the definition of an inferior partner. Suppose further that b is acceptable for a, since $\mu(a) \succ_a b$ immediately follows otherwise. Then, a needs to be matched to $\mu(a) \neq a$, as μ leaves no mutually-acceptable pairs of singles. First, suppose that a and

$\mu(a)$ are matched during Phase 1. Then, $\mu(a)$ is either $\pi(a)$ or $\sigma(a)$, and in either case, $\mu(a) \succ_a b$ follows for b is inferior for a and $b \neq \mu(a)$. Next, suppose that a and $\mu(a)$ are matched at step t of Phase 2. If $x_t = a$, then $\mu(a)$ is the best preferred agent among $\widetilde{\Sigma}_t$. Since $b \in \widetilde{\Sigma}_t$ by assumptions, $\mu(a) \succ_a b$ follows. If $x_t = \mu(a)$, then $\mu(a)$ is superior for a, and hence $\mu(a) \succ_a b$. ■

B Construction of a SaRD Matching up to Depth 3

In this appendix we demonstrate how we can always construct, and thereby guarantee the existence of, a matching that is SaRD up to depth $k = 3$. In B.1, we present a set of sufficient conditions for a matching to be SaRD up to depth $k = 3$. We describe in B.2 an algorithm to compute a matching for an arbitrarily given problem $(N, \succ)$, and then we establish in B.3 that its outcome always satisfies the set of conditions we identify in B.1.

B.1 Sufficient Conditions

In addition to Properties 1–3, we will assume the following property to guarantee a matching to be SaRD up to depth 3.

Property 4. *For any $a \in I^\circ_\mu$, $\sigma^2(a) \notin I^\circ_\mu$.*

The goal of this subsection is to establish the following result:

Proposition 6. *Suppose $\nu \rhd_D \mu$, where μ is a regular matching satisfying Properties 1–4 with respect to a same party permutation σ. Then, (D, ν) is not robust up to depth (at most)* 3.

Before we provide the proof of Proposition 6, we show some further implications of Properties 1 and 2 in B.1.1. Utilizing these implications, along with those in Appendix A.1.1, we prove Proposition 6 in B.1.2.

B.1.1 Further Implications of Properties 1–2

Lemma 6. *Suppose $\nu \rhd_D \mu$, where μ is a regular matching satisfying Property 2. If $a \in Cy$, then $(\pi \circ \nu)^t(a) \in Cy$ for all $t \in \mathbb{N}$.*

Proof. This is an immediate corollary of Lemma 3. ■

Lemma 7. *Suppose $\nu \rhd_D \mu$, where μ is a regular matching satisfying Property 2. If $a \in S_\nu - D$, $\nu(a) \neq \sigma(a)$, and $(\pi \circ \nu)(a) \in S_\nu$, then $(\pi \circ \nu)(a) \in Ch$.*

Proof. Note that $\nu(a) = \mu(a) \in I^\circ_\mu$ follows from $a \in S_\nu - D$ and $\nu(a) \neq \sigma(a)$. By Property 2, $\pi(\nu(a))$ is matched to an inferior agent at μ. For $(\pi \circ \nu)(a) \in S_\nu$ to hold, hence, $(\pi \circ \nu)(a) \in D$ is necessary. Further, $(\pi \circ \nu)(a) \notin Cy$ must follow, because otherwise Lemma 6 entails $a \in Cy \subseteq D$, which contradicts the assumption of $a \notin D$. ■

Lemma 8. *Suppose $\nu \rhd_D \mu$, where μ is a regular matching satisfying Property 2. For any $a \in Cy$, then, $P(a)$ is an odd party.*

Proof. Fix an arbitrary member a of Cy. By definition, $(\pi \circ \nu)^t(a) = a$ for some $t \in \mathbb{N}$. Let $b := (\pi \circ \nu)^{t-1}(a)$, or equivalently $b := \nu(\sigma(a))$, so that b is another member of Cy by Lemma 6. As $b \in D \cap S_\nu$ implies $\nu(b) \in D - S_\nu$, we should have $\nu(b) \in I^\circ_\mu$ and hence, $P(\nu(b))$ is odd. Recalling that $\nu(b) \equiv \sigma(a)$ and hence $P(a) = P(\nu(b))$, the proof is complete. ■

B.1.2 Proof of Proposition 6

To begin, assume that $Ch = \varnothing \neq Cy$. This is without any loss, since otherwise ν is not robust up to depth 1 by Lemma 5. Next fix an agent $b \in \nu(Cy) := \{x \in N | x = \nu(y) \text{ for some } y \in Cy\}$ such that $\sigma^3(b) \notin Cy$. This is again without loss of generality for the following reason: Such b would not exist only if $\sigma^4(b') \in \nu(Cy)$ holds for all $b' \in \nu(Cy)$. This cannot be the case, however, since $b' \in \nu(Cy)$ implies by Lemmas 6 and 8 that $P(b')$ is an odd party.[16] Let $m \in \mathbb{N}$ be such that $|P(b)| = 2m + 1$, and

[16] If P is odd, $|P| \mod 4$ must be 1 or 3. In either case, if $b' \in P \cap \nu(Cy) \Rightarrow \sigma^4(b') \in P \cap \nu(Cy)$, it follows that $P \subset \nu(Cy)$, which is a contradiction.

define $c_j := \sigma^j(b)$ for $j \in \{1, \ldots, 2m\}$. Remember that $\mu(a) \neq a$ by Lemma 2, where $a := \nu(b)$. Therefore, to establish the non-robustness of the original deviation ν up depth κ, it suffices to construct a sequence of κ further deviations such that [1] $\nu_\kappa \rhd_{D_\kappa} \cdots \nu_1 \rhd_{D_1} \nu$, [2] $a \notin D_1 \cup \cdots \cup D_\kappa$, and [3] $b \in D_\kappa$.

If $c_1 \notin S_\nu$, ν is not robust up to depth 1 since we can immediately construct ν_1 by matching b and c_1 so that $\nu_1 \rhd_{\{b,c_1\}} \nu$. For the rest of the proof, thus, we investigate two subcases of $c_1 \in S_\nu$.

Case 1: $c_1 \in S_\nu$ and $\nu(c_1) \neq c_2$. In this case, we can show $c_1 \notin D$ as follows. Suppose towards a contradiction that $c_1 \in D \cap S_\nu$. Since $Ch = \emptyset$ is assumed, c_1 must be another member of Cy. As N is finite, $c_1 \in Cy$ is possible only if $(\pi \circ \nu)^t(c_1) = c_1$ for some $t \in \mathbb{N}$. By Lemma 6, thus, $(\pi \circ \nu)^{t-1}(c_1) \equiv (\pi \circ \nu)^{-1}(c_1)$ is also in $Cy \subseteq (D \cap S_\nu)$. It then follows that $c_2 \in D - S_\nu$ because by definition, $(\pi \circ \nu)^{-1}(c_1) \equiv \nu(\sigma(c_1)) \equiv \nu(c_2)$. However, this contradicts Property 4 as we have both $b \in D - S_\nu$ and $\sigma^2(b) \equiv c_2 \in D - S_\nu$, which respectively imply $b \in I^\circ_\mu$ and $\sigma^2(b) \in I^\circ_\mu$.

As we now have $c_1 \in S_\nu - D$ in addition to the assumptions of $Ch = \emptyset$ and of $\nu(c_1) \neq c_2 \equiv \sigma(c_1)$, Lemma 7 implies $(\pi \circ \nu)(c_1) \notin S_\nu$. We can construct ν_1 and ν_2, respectively by matching $\{\pi(\nu(c_1)), \nu(c_1)\}$ and $\{c_1, b\}$, so that $\nu_2 \rhd_{\{c_1,b\}} \nu_1 \rhd_{\{\pi(\nu(c_1)),\nu(c_1)\}} \nu$. That is, the original deviation ν is not robust up to depth 2.

Case 2: $c_1 \in S_\nu$ and $\nu(c_1) = c_2$. This case arises only when $\mu(c_1) = c_2$, as Property 4 entails $c_2 \notin I^\circ_\mu$. Note further that $|P(b)| \geq 5$ is also necessary; if $|P(b)| = 3$, $c_2 = \pi(b) = (\pi \circ \nu)(a)$ should be a member of $Cy \subseteq S_\nu$, which contradicts $\nu(c_2) = c_1$ being inferior for c_2. That is, $c_3 \equiv \sigma^3(b) \neq b$ should exist in this case. If $c_3 \notin S_\nu$, then ν is not robust up to depth 2, because we can construct ν_1 and ν_2 by respectively matching $\{c_2, c_3\}$ and $\{b, c_1\}$, so that $\nu_2 \rhd_{\{b,c_1\}} \nu_1 \rhd_{\{c_2,c_3\}} \nu$.

For the rest of the proof, we consider the case of $c_3 \in S_\nu$. We then should have $c_3 \notin D$, because the assumptions of $c_3 \equiv \sigma^3(b) \notin Cy$ and $Ch = \emptyset$ entail $c_3 \notin D \cap S_\nu \equiv Cy \cup Ch$. First, suppose $\nu(c_3) = \mu(c_3) \neq c_4$. Then, as in the last part of Case 1, Lemma

7 implies $(\pi \circ \nu)(c_3) \notin S_\nu$. Therefore, we can construct ν_1, ν_2, and ν_3, by respectively matching $\{\pi(\nu(c_3)), \nu(c_3)\}$, $\{c_2, c_3\}$, and $\{b, c_1\}$, so that

$$\nu_3 \rhd_{\{b,c_1\}} \nu_2 \rhd_{\{c_2,c_3\}} \nu_1 \rhd_{\{\pi(\nu(c_3)),\nu(c_3)\}} \nu.$$

That is, the original deviation ν is not robust up to depth 3.

Second, suppose $\nu(c_3) = \mu(c_3) = c_4$. This requires $|P(b)| \geq 7$, since if $|P(b)| = 5$, the original assumption of $b \in \nu(Cy)$ implies $c_4 = \pi(b) \in Cy$, which is incompatible with $\nu(c_3) = c_4$. Then, $c_5 \in S_\nu$ cannot hold for the following reason:

- If $|P(b)| = 7$, the original assumption of $b \in \nu(Cy)$ implies $c_6 = \pi(b) \in Cy \subseteq D$. Then $c_5 \in S_\nu$ would require $c_5 \in D$ and hence $c_5 \in Cy$, as $\mu(c_5) = c_6$ by Property 3 and $Ch = \emptyset$ by assumption. By the definition of Cy, however, $c_5 \in Cy$ implies $c_6 \equiv \sigma(c_5) \notin S_\nu$, which is incompatible with $c_6 \in Cy$.
- If $P(b) > 7$, since $c_5 \in I^\circ_\mu$ by Property 3, $c_5 \in S_\nu$ would again require $c_5 \in D$, which is followed by $c_5 \in Cy$ and $c_6 \in \nu(Cy)$. This is a contradiction, because Property 3 implies $c_6 \notin I^\circ_\mu$ while $\nu(Cy) \subseteq D - S_\nu$ by definition.

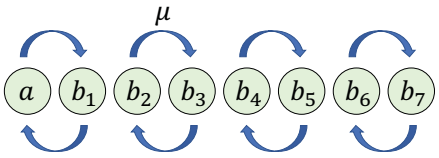
Given $c_5 \notin S_\nu$, we can construct ν_1, ν_2, and ν_3, by respectively matching $\{c_4, c_5\}$, $\{c_2, c_3\}$, and $\{b, c_1\}$, so that

$$\nu_3 \rhd_{\{b,c_1\}} \nu_2 \rhd_{\{c_2,c_3\}} \nu_1 \rhd_{\{c_4,c_5\}} \nu.$$

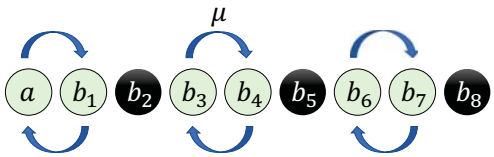
That is, the original deviation ν is not robust up to depth 3. ■

B.2 Description of Algorithm B

Taking a problem $(N, \succ)$ and a party permutation σ as given, construct a matching μ as follows. To simplify the description, we write "define $\mu(a) := b$," when it should read as "define $\mu(a) := b$ and $\mu(b) := a$." The whole procedure is divided into five phases.



(a) Phase 1: $P \in \mathscr{E}$



(b) Phase 2: $P \in \mathscr{O}_{3\times}$

Figure 2: Matching during Phases 1 and 2 of the Algorithm. For each j, b_j represents $\sigma^j(a)$. Each arrow between two agents means they are matched, and the agents represented by black circles are not matched in Phase 2.

B.2.1 Phase 1 of Algorithm B

Let $\mathscr{E} \subseteq \mathscr{P}(\sigma)$ be the family of even parties; i.e., $\mathscr{E} := \{P \in \mathscr{P}(\sigma) : |P| \text{ is even}\}$. For each $E \in \mathscr{E}$, arbitrarily take $a \in E$ and define $\mu\left(\sigma^{2j}(a)\right) = \sigma^{2j+1}(a)$ for each $j \in \left\{1, \ldots, \frac{|P|}{2}\right\}$, as illustrated in Figure 2 (a).

B.2.2 Phase 2 of Algorithm B

Let $\mathscr{O}_{3\times} \subseteq \mathscr{P}(\sigma) - \mathscr{E}$ be the family of odd parties whose sizes are a multiple of three; i.e., $\mathscr{O}_{3\times} := \{P \in \mathscr{P}(\sigma) - \mathscr{E} : |P| = 3n \text{ for some } n \in \mathbb{N}\}$. For each $P \in \mathscr{O}_{3\times}$, arbitrarily take $a \in P$ and define $\mu\left(\sigma^{3j}(a)\right) = \sigma^{3j+1}(a)$ for each $j \in \left\{1, \ldots, \frac{|P|}{3}\right\}$, as illustrated in Figure 2 (b).

Remark 1. *Phases 1 and 2 simply match "adjacent" pairs of agents (with respect to σ) within each party, as illustrated in Figure 2. Note that every member of each $P \in \mathscr{E}$ is matched in Phase 1, while there are $|P|/3$ unmatched agents in each $P \in \mathscr{O}_{3\times}$ in Phase 2. Note also that if $P(a) \in \mathscr{O}_{3\times}$ and a is not matched in this phase, then $\pi(a)$ and $\sigma(a)$ are matched, respectively, to $\pi^2(a)$ and $\sigma^2(a)$.* □

B.2.3 Phase 3 of Algorithm B

Let $U_0 \subseteq N$ be the set of agents who are not matched yet and $\mathscr{U}_0 := \mathscr{P}(\sigma) - (\mathscr{E} \cup \mathscr{O}_{3\times})$ be the family of parties none from which is matched yet.[17] Arbitrarily order the members of U_0 as $x_1, \ldots, x_{|U_0|}$ and iterate the following step for $t = 1, \ldots, |U_0|$.

Remark 2. *In what follows, U_t and $\mathscr{U}_t$ will be, respectively, the set of agents who are unmatched by step t and the family of parties no agent from which is matched by step t.* □

Step $t = 1, \ldots, |U_0|$ of Phase 3:

If $x_t \notin U_{t-1}$, then, proceed to step $t+1$ with $U_t = U_{t-1}$ and $\mathscr{U}_t = \mathscr{U}_{t-1}$. Otherwise, define

$$\Sigma_t := \left\{ y \in U_{t-1} - \{\pi(x_t), \pi^2(x_t)\} : x_t \text{ is superior for } y \text{ and } y \text{ is acceptable for } x_t \right\}.$$

If Σ_t is empty, then proceed to step $t+1$ with $U_t = U_{t-1}$ and $\mathscr{U}_t = \mathscr{U}_{t-1}$.[18] Otherwise, let $y_t \in \Sigma_t$ denote the best partner for x_t among those in Σ_t; that is, $y \in \Sigma_t \Rightarrow y_t \succeq_{x_t} y$. Define $\mu(x_t) = y_t$ and $\mathscr{U}_t = \mathscr{U}_{t-1} - \{P(x_t), P(y_t)\}$. If $\mathscr{U}_t = \mathscr{U}_{t-1}$, proceed to step $t+1$ with $U_t = U_{t-1} - \{x_t, y_t\}$. Otherwise, further divide the case as follows.

Case 1: $P(x_t) = P(y_t) \in \mathscr{U}_{t-1}$.

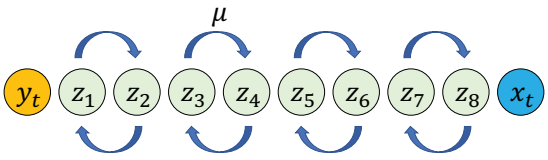
In this case, there exist $q, r \in \{1, \ldots, |P(x_t)|\}$ such that $\sigma^{q+1}(y_t) = x_t$ and $\sigma^{r+1}(x_t) = y_t$. Notice that one and only one of them is odd, for $|P(x_t)| = q + r + 2$ must be odd by definition. It should be also noted that $q \geq 2$ by the definition of Σ_t. Match the agents in $P(x_t) = P(y_t)$ as follows:

- Matching among $\sigma(y_t), \ldots, \sigma^q(y_t)$:

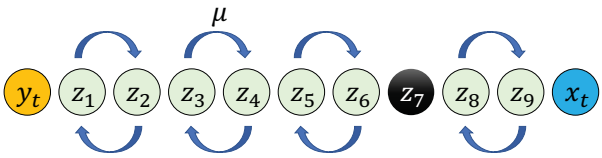
  If $q = 2m$ for some $m \in \mathbb{N}$, then $\mu\left(\sigma^{2j-1}(y_t)\right) = \sigma^{2j}(y_t)$ for each $j \in \{1, \ldots, m\}$. If $q = 2m+1$ for some $m \in \mathbb{N}$, then $\mu\left(\sigma^{2j-1}(y_t)\right) = \sigma^{2j}(y_t)$ for each $j \in \{1, \ldots, m -$

[17] Remember that $a \in U_0$ does not necessarily imply $P(a) \in \mathscr{U}_0$.

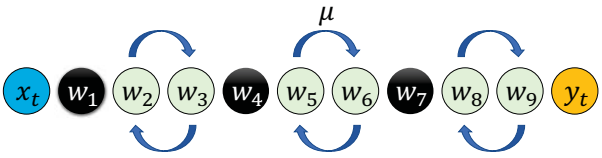
[18] Remember that when $\{x_t\} \in \mathscr{P}(\sigma)$ is a solitary party, y is acceptable for x_t if and only if y is superior for x_t, which can be the case only if x_t is inferior for y. In such a case, thus, Σ_t must be empty.



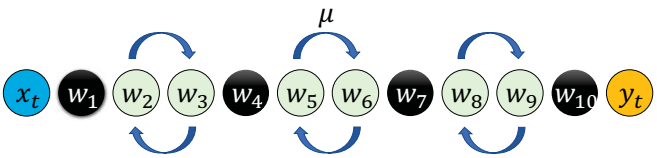
(a) Case of q being even



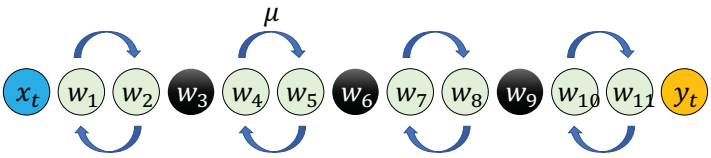
(b) Case of q being odd



(c) Case of $r = 3n$ for some $n \in \mathbb{N}$

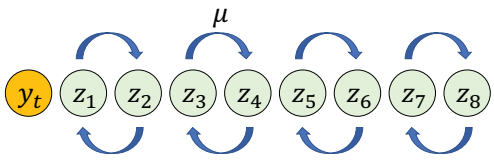


(d) Case of $r = 3n + 1$ for some $n \in \mathbb{N} \cup \{0\}$

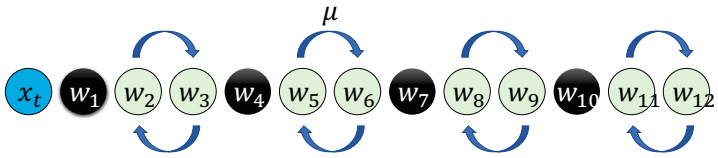


(e) Case of $r = 3n + 2$ for some $n \in \mathbb{N} \cup \{0\}$

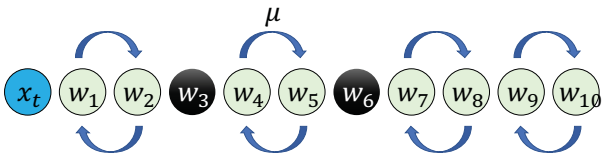
Figure 3: Matching in Case 1 of Phase 3. For each j, z_j and w_j denote $\sigma^j(y_t)$ and $\sigma^j(x_t)$, respectively. Each arrow between two agents means they are matched, and the agents represented by black circles are not matched in this step.



(a) Matching of $P(y_t)$



(b) Matching of $P(x_t)$ with $|P(x_t)| = 3n + 1$ for some $n \in \mathbb{N}$



(c) Matching of $P(x_t)$ with $|P(x_t)| = 3n + 2$ for some $n \in \mathbb{N}$

Figure 4: Matching of the agents in $P(x_t), P(y_t) \in \mathscr{U}_{t-1}$ in Case 2 of Phase 3. For each j, z_j and w_j denote, respectively, $\sigma^j(y_t)$ and $\sigma^j(x_t)$. Each arrow between two agents means they are matched, and the agents represented by black circles are not matched in this step.

$1\}$, and $\mu(\left(\sigma^{2m}(y_t)\right) = \sigma^{2m+1}(y_t)$, leaving $\mu\left(\sigma^{2m-1}(y_t)\right)$ undefined. Figure 3 (a)–(b) illustrate the matching in these cases.

- Matching among $\sigma(x_t), \ldots, \sigma^r(x_t)$:

  If $r = 3n$ or $3n + 1$ for some $n \in \mathbb{N} \cup \{0\}$, then, let $\mu\left(\sigma^{3j'-1}(x_t)\right) = \sigma^{3j'}(x_t)$ for each $j' \in \{1, \ldots, n\}$. Notice that $\mu\left(\sigma^{3n+1}(x_t)\right)$ is undefined when $r = 3n + 1$.

  If $r = 3n + 2$ for some $n \in \mathbb{N} \cup \{0\}$, then, let $\mu\left(\sigma^{3j'-2}(x_t)\right) = \sigma^{3j'-1}$ for each $j' \in \{1, \ldots, n + 1\}$. Figure 3 (c)–(e) illustrate the matching in these cases.

Let $U_t := U_{t-1} - M_t$, where M_t is the set of agents matched in this step, including x_t and y_t, and proceed to step $t + 1$.

Case 2: $P(x_t) \neq P(y_t)$.

In this case, match the members of $P(x_t)$ and $P(y_t)$, respectively, if $P(x_t) \in \mathscr{U}_{t-1}$ and $P(y_t) \in \mathscr{U}_{t-1}$ as follows:

- Matching among $P(y_t) \in \mathscr{U}_{t-1}$:
  If $P(y_t) \in \mathscr{U}_{t-1}$, define $\mu\left(\sigma^{2j-1}(y_t)\right) := \sigma^{2j}(y_t)$ for each $j = 1, \ldots, \frac{|P(y_t)|-1}{2}$, as illustrated in Figure 4 (a).
- Matching among $P(x_t) \in \mathscr{U}_{t-1}$:[19]
  If $P(x_t) \in \mathscr{U}_{t-1}$ and $|P(x_t)| = 3n+1$ for some $n \in \mathbb{N}$, then, let $\mu\left(\sigma^{3j'-1}(x_t)\right) = \sigma^{3j'}(x_t)$ for each $j' \in \{1, \ldots, n\}$. If $P(x_t) \in \mathscr{U}_{t-1}$ and $|P(x_t)| = 3n+2$ for some $n \in \mathbb{N}$, then let $\mu\left(\sigma^{3j'-2}(x_t)\right) = \sigma^{3j'-1}(x_t)$ for each $j' \in \{1, \ldots, n\}$ and $\mu\left(\sigma^{3n}(x_t)\right) = \sigma^{3n+1}(x_t)$. Figures 4 (b)–(c) illustrate the matching in these cases.

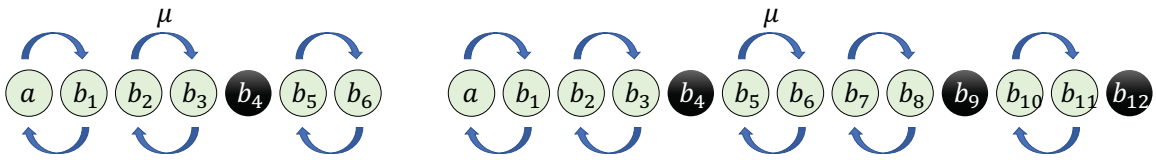
Let $U_t := U_{t-1} - M_t$, where M_t is the set of agents matched in this step, including x_t and y_t, and proceed to step $t+1$.

Remark 3. *To see the point in this phase, suppose that $x_{t'}$ is matched to $y_{t'}$ in step t' of this phase.*

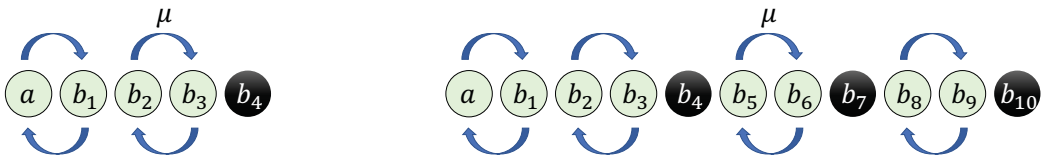
- *If $P(x_{t'}) \in \mathscr{U}_{t'-1}$, then $\sigma^2(x_{t'})$ is matched to either $\sigma(x_{t'})$ or $\sigma^3(x_{t'})$, and $\pi(x_{t'})$ is always matched to $\pi^2(x_{t'})$.*
- *If $P(x_{t'}) \notin \mathscr{U}_{t'-1}$, then $\sigma^2(x_{t'})$ is again matched to either $\sigma(x_{t'})$ or $\sigma^3(x_{t'})$, and we have $\mu(\pi(x_{t'})) \neq \pi^2(x_{t'})$ only if $x_t = \pi(x_{t'})$ is matched to some y_t an earlier step $t < t'$ such that $P(x_t) \in \mathscr{U}_{t-1}$.*[20] □

Remark 4. *For any $x_t, x_{t'} \in U_{|U_0|}$, we have either (i) they are not mutually acceptable, (ii) they are mutually inferior to each other, or (iii) $P(x_t) = P(x_{t'}) \in \mathscr{U}_{|U_0|}$. To see this, suppose that $x_t, x_{t'} \in U_{|U_0|}$ are mutually acceptable and that x_t is superior for $x_{t'}$. For x_t to be not matched in step t of Phase 3, then, we should have $\Sigma_t \not\ni x_{t'}$. By the definition of Σ_t, it entails $x_{t'} \in \{\pi(x_t), \pi^2(x_{t'})\}$ and thus $P(x_t) = P(x_{t'}) \in \mathscr{U}_{|U_0|}$.* □

[19] As $\mathscr{U}_{t-1} \cap \mathscr{O}_{3\times} = \emptyset$ by definition, $P(x_t) \in \mathscr{U}_{t-1}$ implies that $|P(x_t)|$ is not a multiple of three.
[20] For instance, take $x_{t'}$ to be w_1 in Figure 4 (b).



(a) $P \in \mathscr{V}$ with $|P| = 7$ (b) $P \in \mathscr{V}$ with $|P| = 3n+1 > 7$ for some $n \in \mathbb{N}$



(c) $P \in \mathscr{V}$ with $|P| = 5$ (d) $P \in \mathscr{V}$ with $|P| = 3n+2$ for some $n \in \mathbb{N}$

Figure 5: Matching during Phase 4. For each j, b_j denotes $\sigma^j(a)$. Each arrow between two agents means they are matched, and the agents represented by black circles are not matched in this Phase.

B.2.4 Phase 4 of Algorithm B

Let $\mathscr{V} := \mathscr{U}_{|U_0|}$, i.e., the family of odd parties no member from which has been matched yet. For each $P \in \mathscr{V}$, fix an arbitrary member $a \in P$ and match the members of P in the following way, as illustrated in Figure 5:

- If $|P| = 3n+1$ for some $n \in \mathbb{N}$, then, define $\mu(a) := \sigma(a)$, $\mu\left(\sigma^2(a)\right) := \sigma^3(a)$, $\mu\left(\sigma^5(a)\right) := \sigma^6(a)$, and $\mu\left(\sigma^{3j-2}(a)\right) := \sigma^{3j-1}(a)$ for each $j \in \{3, \ldots, n\}$.[21]
- If $|P| = 3n+2$ for some $n \in \mathbb{N}$, then define $\mu(a) := \sigma(a)$, $\mu\left(\sigma^2(a)\right) := \sigma^3(a)$, and $\mu\left(\sigma^{3j-1}(a)\right) := \sigma^{3j}(a)$ for each $j \in \{2, \ldots, n\}$.

Remark 5. *As illustrated in Figure 5, if $P(\alpha) \in \mathscr{V}$ but $\alpha \in N$ is not matched in this phase, $\pi(\alpha)$ and $\sigma(\alpha)$ are matched, respectively, to $\pi^2(\alpha)$ and $\sigma^2(\alpha)$. Combined with Remarks 1 and 3, if $\alpha \in I_\mu^\circ$ at the final outcome,*

- *$\sigma^2(\alpha)$ is matched to either $\sigma(\alpha)$ or $\sigma^3(\alpha)$, and*
- *$\pi(\alpha)$ is matched to $\pi^2(\alpha)$, unless $\pi(\alpha) = x_t$ is matched to y_t in step t such that $P(\alpha) \in \mathscr{U}_{t-1}$ during Phase 3.* □

[21] As P is an odd party, $|P| = 3n+1$ for some $n \in \mathbb{N}$ implies $|P| \geq 7$.

Remark 6. *If agents α and β both remain unmatched by the end of this phase, they are either (i) not mutually acceptable or (ii) mutually inferior to each other. At the end of Phase 3 they have a third possibility, $\alpha \in \{\pi(\beta), \pi^2(\beta)\}$ or $\beta \in \{\pi(\alpha), \pi^2(\alpha)\}$, as argued in Remark 4, but not both of such α and β can remain unmatched after Phase 4 matches the agents in $P(a) = P(\beta)$ as specified above.* □

B.2.5 Phase 5 of Algorithm B

Let R_0 be the set of those who still remain unmatched, and arbitrarily order its members as $r_1, \ldots, r_{|R_0|}$. Iterate the following step for $\tau = 1, \ldots, |R_0| + 1$:

Step $\tau = 1, \ldots, |R_0|$ of Phase 5:
If $r_\tau \in R_{\tau-1}$ and there exists some $r_i \in R_{\tau-1}$ who is mutually acceptable with r_τ, then define $\mu(r_\tau) := r_i$ and proceed to step $\tau + 1$ with $R_\tau := R_\tau - \{r_\tau, r_i\}$.[22] Otherwise, proceed to step $\tau + 1$ with $R_\tau := R_{\tau-1}$.

Step $|R_0| + 1$ of Phase 5:
For any $r \in R_{|R_0|}$, i.e., for any agent not matched yet, define $\mu(r) = r$.

B.3 Properties of Algorithm B

As mentioned at the beginning of this Section, the above algorithm is designed so that its outcomes always satisfy Properties 1–4. Here we formally establish this fact.

Proposition 7. *Let μ be an outcome of the algorithm of Section B.2. Then, it is regular and satisfies Properties 1–4 (with respect to the party permutation σ fixed at the beginning of the algorithm).*

Proof of regularity. It is immediate to check that μ is individually rational as we only

[22] In general multiple members of $R_{\tau-1}$ may be mutually acceptable with r_τ. Even if so, the choice of r_i can be arbitrary.

match mutually-acceptable pairs during the algorithm, and it leaves no mutually-acceptable pairs of singles because of Phase 5. ■

Proof of Property 1. Suppose that a is superior for b and $\mu(b) = b$, where μ is an outcome of the algorithm. Also assume that b is acceptable for a, as otherwise $\mu(a) \succ_a b$ immediately follows from individual rationality. Then, a should be matched to $\mu(a)$ by the end of Phase 4; otherwise, the assumptions are incompatible as argued in Remark 6. If $\mu(a) = \pi(a)$, then $\mu(a) = \pi(a) \succ_a b$ immediately follows, since our assumptions of a being superior for b and of $\mu(b) = b$ respectively imply that b is inferior for a and $b \neq \mu(a) = \pi(a)$. If $\mu(a) = \sigma(a) \neq \pi(a)$, then we also obtain $\mu(a) \succ_a \pi(a) \succeq_a b$ by the definition of a (semi-)party permutation.

What remains to check is the case where a is matched to $\mu(a) \notin \{\pi(a), \sigma(a)\}$ during Phase 3. If $a = y_t$ is matched to x_t in some step t during Phase 3, $\mu(a) = x_t$ is superior for $a = y_t$ and hence, $\mu(a) \succ_a b$ holds. If $a = x_t$ is matched to y_t in some step t during Phase 3, our assumptions imply $b \in \Sigma_t$.[23] Therefore, $\mu(a) \succ_a b$ holds by the definition of y_t. ■

Proof of Property 2. Suppose $a \in I^{\circ}_{\mu}$, where μ is an outcome of the algorithm. As this implies $\mu(a) \notin \{\pi(a), \sigma(a)\}$, it is immediate to see that $P(a)$ is odd; otherwise, a and $\mu(a) \in \{\pi(a), \sigma(a)\}$ should be matched during Phase 1. Moreover, by the arguments in Remark 5, either $\mu(\pi(a)) = \pi^2(a)$ or $\pi(a) = x_t$ is matched to y_t in some step t during Phase 3. In either case, $\mu(\pi(a))$ is inferior for $\pi(a)$. ■

Proof of Property 3. Suppose $a \in I^{\circ}_{\mu}$, $\mu(\sigma(a)) = \sigma^2(a)$ and $\mu\left(\sigma^3(a)\right) = \sigma^4(a)$, where μ is an outcome of the algorithm, and also $|P(a)| \geq 7$ since the claim vacuously holds otherwise. Note that $P(a) \in \mathscr{U}_0$, because $P(a) \in \mathscr{O}_{3\times}$ is incompatible with the assumptions. Therefore, if $P(a) \notin \mathscr{V}$, there exists some t such that $P(a) \in \mathscr{U}_{t-1} - \mathscr{U}_t$. For the

[23] In this case, $b \notin \{\pi(a), \pi^2(a)\}$ holds for the following reason: As we assume $\mu(b) = b$, it suffices to confirm that neither $\pi(a)$ nor $\pi^2(a)$ is single at μ, which is clearly true if $\mu(\pi(a)) = \pi^2(a)$. Given $a = x_t$ is matched to y_t during Phase 3, $\mu(\pi(a)) = \pi^2(a)$ fails only if $\pi(a) = x_{t'}$ is matched to $y_{t'}$ in an earlier step $t' < t$, as argued in Remark 3. Moreover, for both a and $\pi(a)$ to remain unmatched until step t', we must have $P(a) \in \mathscr{U}_{t'-1}$ and hence, $\pi^2(a)$ should be also matched (to $\pi^3(a)$) in step t'.

assumptions of $a \in I^\circ_\mu$, $\mu(\sigma(a)) = \sigma^2(a)$ and $\mu\left(\sigma^3(a)\right) = \sigma^4(a)$ to simultaneously hold, then, the only possibility is that $|P(a)| = 3n + 2$ and $x_t = \sigma^5(a)$, as seen in Figure 4 (c). In such a case, $\mu\left(\sigma^5(a)\right) = y_t$ is inferior for $\sigma^5(a) = x_t$ by definition, and $\sigma^6(a) = \sigma(x_t)$ is matched to $\sigma^7(a) = \sigma^2(x_t)$. That is, we have both $\sigma^5(a) \in I^\circ_\mu$ and $\sigma^6(a) \notin I^\circ_\mu$.

Next, consider the case of $P(a) \in \mathscr{V}$, i.e., the case where none from $P(a)$ is matched by the end of Phase 3. If $P(a) \in \mathscr{V}$ and $|P(a)| = 7$, then $\mu\left(\sigma^5(a)\right) = \sigma^6(a)$ as shown in Figure 5 (a). If $P(a) \in \mathscr{V}$ and $|P(a)| > 7$, then, $\sigma^5(a)$ is not matched during Phase 4 and $\sigma^6(a)$ is matched to $\sigma^7(a)$, as illustrated in Figure 5 (b)–(d). Since $\sigma^5(a)$ cannot match to a superior parter during Phase 5 as argued in Remark 6, these imply $\sigma^5(a) \in I^\circ_\mu$ and $\sigma^6(a) \notin I^\circ_\mu$ as required. ■

Proof of Property 4. Suppose $a \in I^\circ_\mu$, where μ is an outcome of the algorithm. As argued in Remark 5, then, $\sigma^2(a)$ should be matched to $\sigma(a)$ or $\sigma^3(a)$ and in either case, $\sigma^2(a) \notin I^\circ_\mu$ holds. ■

C Other Proofs

C.1 Proof of Proposition 1

If μ is not individually rational, i.e., if $a \succ_a \mu(a)$ for some a, then the deviation $(\{a\}, \nu)$ is robust up to any depth $k \geq 1$, where $\nu(b) = b$ if $b \in \{a, \mu(a)\}$ and $\nu(b) = \mu(b)$ otherwise. If μ leaves a mutually-acceptable pair of singles, i.e., if $a \succ_b b$, $b \succ_a a$, $\mu(a) = a$ and $\mu(b) = b$ for some a and b, then, the deviation $(\{a, b\}, \nu')$ is robust up to any depth $k \geq 1$, where $\nu'(a) = b$, $\nu'(b) = a$, and $\nu(c) = \mu(c)$ for all $c \in N - \{a, b\}$. ■

C.2 Proof of Proposition 2

The proof is by examples. First, we provide an example of a matching that is SaRD up to depth 1 (and hence, up to any depth $k \geq 1$) but not in the bargaining set. Let $N = \{1, 2, 3\}$ and $\succ_i$ be such that $(i + 1) \succ_i (i - 1) \succ_i i \pmod 3$ for each $i \in N$. In

this problem, it is easy to check that $\mu = \{\{1,2\},\{3\}\}$ is SaRD up to depth 1: $(D,\nu) = (\{2,3\},\{\{1\},\{2,3\}\})$ is the only deviation from μ, and this is not robust as $\nu' \rhd_{\{1,3\}} \nu$ and agent $2 \in D$ gets strictly worse off at ν' than at μ, where $\nu' = \{\{1,3\},\{2\}\}$. However, this μ is not in the bargaining set, because $\nu'(1) = 3 \not\succeq_1 2 = \mu(1)$ and hence, $(\{1,3\},\nu')$ is not qualified to be a counterobjection against $(\{2,3\},\nu)$.

The second is an example of a matching that is in the bargaining set but not SaRD up to any depth $k \geq 1$. Let $N = \{m_1, m_2, w_1, w_2, w_3\}$ and $\succ$ be such that

$$w_1 \succ_{m_1} w_2 \succ_{m_1} w_3 \succ_{m_1} m_1 \succ_{m_1} m_2, \quad w_2 \succ_{m_2} w_1 \succ_{m_2} w_3 \succ_{m_2} m_2 \succ_{m_2} m_1,$$
$$m_2 \succ_{w_1} m_1 \succ_{w_1} w_1 \succ_{w_1} w_2 \succ_{w_1} w_3, \quad m_1 \succ_{w_2} m_2 \succ_{w_2} w_2 \succ_{w_2} w_1 \succ_{w_2} w_3, \quad \text{and}$$
$$w_3 \succ_{w_3} m_1 \succ_{w_3} m_2 \succ_{w_3} w_1 \succ_{w_3} w_2.$$

This problem is a marriage problem with $M = \{m_1, m_2\}$ and $W = \{w_1, w_2, w_3\}$. It is easy to verify that $\mu = \{\{m_1\},\{m_2\},\{w_1\},\{w_2\},\{w_3\}\}$ is in Zhou's (1994) bargaining set defined in Section 4.1.1. However, Proposition 1 implies that this μ is not SaRD up to any depth k, as it leaves mutually-acceptable pairs of singles. ■

C.3 Proof of Proposition 3

For each $\mathscr{P}$-stable matching μ', by appropriately choosing a_P's in Phase 1, we can take an outcome μ of Algorithm A in Appendix A.2 so that μ includes μ'. Thus, the claim of the proposition immediately follows from Propositions 4–5 in Appendix A. ■

D Tightness of Theorems 1–2 with respect to σ

In this appendix we demonstrate that the sufficient conditions in Theorems 1–2 are (almost) tight among those depending only on party permutations. The results here should suggest that a tighter sufficient condition would be significantly more complicated, since it would require detailed information of a preference profile. First, we can

show that the condition in Theorem 1 is tight in the following sense:

Proposition 8. *Let σ be a permutation over N such that $|P| = 2m+1$ for some $P \in \mathscr{P}(\sigma)$ where $m \geq 2$. Then, there exists $\succ = (\succ_i)_{i \in N}$ such that σ is a party permutation for $(N, \succ)$ and no matching is SaRD up to depth 1 in $(N, \succ)$.*

Proof. Given σ, consider a preference profile such that for each $a \in N$, (i) only $\sigma(a)$ and $\pi(a)$ are acceptable and (ii) $\sigma(a) \succ_a \pi(a)$ if $|P(a)| > 1$. Then, σ is a party permutation for $(N, \succ)$.

Fix an arbitrary regular matching μ for $(N, \succ)$ and an odd party $P \in \mathscr{P}(\sigma)$ with $|P| \geq 5$. By the regularity of μ, there must be a such that $\mu(a) = a$, $\mu(\pi(a)) = \pi^2(a)$, and $\mu(\sigma(a)) = \sigma^2(a)$. Let (D, ν) be the deviation by $D = \{a, \pi(a)\}$ from μ. Then, it is robust up to depth 1 for the following reason: Since $\pi(a)$ is matched to her best possible partner, a, she would not deviate from ν. As $\sigma(a)$ is the only partner whom a prefers to $\pi(a)$, $\sigma(a) \in D'$ is necessary for (D', ν') to be a deviation from ν with $a \in D'$. However, $|P| > 3$ implies $\sigma^2(a) \notin D$ and hence $\nu(\sigma(a)) = \mu(\sigma(a)) = \sigma^2(a)$; that is, $\sigma(a)$ would not agree to deviate with a from ν. In conclusion, we must have $D \cap D' = \varnothing$ for any deviation (D', ν') from ν, and thus, (D, ν) is robust up to depth 1. ■

The condition in Theorem 2 is "almost" tight in a similar sense. More specifically, it is tight except for special cases where every non-solitary odd party has cardinality 9 or 15.

Proposition 9. *Let σ be a permutation over N such that $|P| = 2m+1$ for some $P \in \mathscr{P}(\sigma)$ where $m \geq 3$ and $m \neq 4, 7$. Then, there exists $\succ = (\succ_i)_{i \in N}$ such that σ is a party permutation for $(N, \succ)$ and no matching is SaRD up to depth 2 in $(N, \succ)$.*

Proof. Given σ, arbitrarily take $P \in \mathscr{P}(\sigma)$ such that $|P| = 2m+1$ for some $P \in \mathscr{P}(\sigma)$ where $m \geq 3$ and $m \neq 4, 7$. We consider two cases separately.

Case 1: $|P| \not\equiv 0 \pmod 3$. In this case, consider a preference profile such that for each $a \in N$, (i) only $\sigma(a)$ and $\pi(a)$ are acceptable and (ii) $\sigma(a) \succ_a \pi(a)$ if $|P(a)| > 1$. Then, σ is a party permutation for $(N, \succ)$. Fix an arbitrary regular matching μ. By regularity and the assumption of $|P| \not\equiv 0 \pmod 3$, there must exist $a \in P$ such that $\mu(a) = a$, $\mu(\pi(a)) = \pi^2(a)$, $\mu(\sigma(a)) = \sigma^2(a)$, and $\mu(\sigma^3(a)) = \sigma^4(a)$.[24] Now consider the deviation (D, ν) from μ by $D = \{a, \pi(a)\}$. By similar arguments as in the proof of Proposition 8, along with the assumption of $|P| \geq 7$, we can confirm the following: for (D_1, ν_1) and (D_2, ν_2) such that $\nu_2 \rhd_{D_2} \nu_1 \rhd_{D_1} \nu$, neither D_1 and D_2 contains a and hence $\nu_1, \nu_2 \succeq_D \mu$.[25] That is, the deviation by $D = \{a, \pi(a)\}$ is robust up to depth 2.

Case 2: $|P| = 3n$ for some odd $n \geq 7$. In this case, first fix an arbitrary agent $a_0 \in P$ and label the agents in P by $a_j = \sigma^j(a_0)$ for each $j \in \mathbb{Z}$. Consider a preference profile $\succ$ such that

- for each a_j with $j \in \{0, 7, 14\}$, (i) only a_{j-1}, a_{j+1} and a_{j+6} are acceptable for $\succ_{a_j}$ and (ii) $a_{j+1} \succ_{a_j} a_{j-1} \succ_{a_j} a_{j+6}$,
- for each a_i with $i \in \{6, 13, 20\}$, (i) only a_{i-1}, a_{i+1} and a_{i-6} are acceptable for $\succ_{a_i}$ and (ii) $a_{i+1} \succ_{a_i} a_{i-6} \succ_{a_i} a_{i-1}$, and
- for any other agent b, (i) only $\sigma(b)$ and $\pi(b)$ are acceptable for $\succ_b$ and (ii) $\sigma(b) \succ_b \pi(b)$.

Note that there are three "remote" mutually-acceptable pairs: $\{a_0, a_6\}$, $\{a_7, a_{13}\}$, and $\{a_{14}, a_{20}\}$. Nonetheless, it is easy to check that σ is a party permutation for $(N, \succ)$. Lastly, arbitrarily fix a regular matching μ for $(N, \succ)$. We consider two subcases.

First, suppose that $\mu(a_{j^*}) = a_{j^*+6}$ for some $a_{j^*} \in P$ with $j^* \in \{0, 7, 14\}$. Then, the regularity of μ implies either

[1] $\mu(a_{j^*+1}) = a_{j^*+2}$, $\mu(a_{j^*+3}) = a_{j^*+4}$, and $\mu(a_{j^*+5}) = a_{j^*+5}$;

[24] The oddness of P, along with regularity, implies the existence of a such that $\mu(a) = a$. Regularity then implies $\mu(\pi(a)) = \pi^2(a)$ and $\mu(\sigma(a)) = \sigma^2(a)$; otherwise, $\{a, \pi(a)\}$ or $\{a, \sigma(a)\}$ is a mutually-acceptable pair of singles. Lastly, if $\mu(a) = a$ and $\mu(\sigma(a)) = \sigma^2(a)$ were to imply $\mu(\sigma^3(a)) \neq \sigma^4(a)$ (and hence $\mu(\sigma^3(a)) = \sigma^3(a)$), it would contradict $|P| \not\equiv 0 \pmod 3$.

[25] Note that we need $|P| \geq 7$ here to guarantee $\pi(a) \neq \sigma^4(a)$ and thus that $\sigma^3(a)$ remains matched to $\sigma^4(a)$ at ν.

[2] $\mu(a_{j^*+1}) = a_{j^*+1}$, $\mu(a_{j^*+2}) = a_{j^*+3}$, and $\mu(a_{j^*+4}) = a_{j^*+5}$; or

[3] $\mu(a_{j^*+1}) = a_{j^*+2}$, $\mu(a_{j^*+3}) = a_{j^*+3}$, and $\mu(a_{j^*+4}) = a_{j^*+5}$.

In the first case, the deviation by $D = \{a_{j^*}, a_{j^*-1}\}$ is robust up to 2 for a similar reason as in the previous case. In the second and third cases, respectively, those by $D = \{a_{j^*}, a_{j^*+1}\}$ and $D = \{a_{j^*+2}, a_{j^*+3}\}$ are robust up to depth 2.[26] In any case, thus, the prefixed matching μ is not robust up to depth 2.

Second, suppose that $\mu(a) \in \{a, \pi(a), \sigma(a)\}$ for all $a \in P$ (and hence, for all $a \in N$). Note that in this case, too, there must exist $a_{i^*+1} \in P$ such that $\mu(a_{i^*+1}) = a_{i^*+2}$ and $\mu(a_{i^*+3}) = a_{i^*+4}$.[27] It is then without any loss to further assume a_{i^*} is single at μ.[28] Given a_{i^*} is single, moreover, a_{i^*-1} needs to be matched to a_{i^*-2} by regularity and the supposition of $\mu(a_{i^*}) \in \{a_{i^*-1}, a_{i^*}, a_{i^*+1}\}$. In sum, we have assumed that μ induces a partition that includes

$$\{a_{i^*-2}, a_{i^*-1}\}, \{a_{i^*}\}, \{a_{i^*+1}, a_{i^*+2}\}, \text{ and } \{a_{i^*+3}, a_{i^*+4}\}.$$

Now we consider the deviation (D, ν) from μ by $D = \{a_{i^*-1}, a_{i^*}\}$ and establish that it is robust up to depth 2. To do so, first suppose $\nu_1 \rhd_{D_1} \nu$. Since a_{i^*-1} and a_{i^*+1} are matched to their best possible parter at ν, they cannot be a member of D_1. When $i^* \in \{6, 13, 20\}$, a_{i^*} prefers a_{i^*-6} to a_{i^*-1}, but a_{i^*+6} should be matched to either a_{i^*-5} or a_{i^*-7} at ν and thus would not agree to deviate with a_{i^*}.[29] Therefore, a_{i^*} cannot be a member of D_1, either. That is, $\{a_{i^*-1}, a_{i^*+1}\}$ remains matched at any ν_1 such that $\nu_1 \rhd_{D_1} \nu$, which establishes the robustness of the original deviation up to depth 1.

[26] To see the robustness of $D = \{a_{j^*+2}, a_{j^*+3}\}$ in the third case, note that a_{j^*+6} is matched to her superior partner a_{j^*} at ν. This makes it impossible for $\{a_{j^*+5}, a_{j^*+6}\}$ to deviate from ν.

[27] To see this suppose otherwise, i.e., suppose that $\mu(a_i + 1) = a_{i+2}$ implies $\mu(a_{i+3}) = a_{i+3}$) for all $i \in \mathbb{Z}$. Then, by the assumption of $|P|$ being a multiple of 3, either of $\{a_0, a_6\}$, $\{a_7, a_{13}\}$, or $\{a_{14}, a_{20}\}$ should be a mutually-acceptable pair of single, which contradicts the regularity of μ.

[28] Otherwise, we can redefine $i_1^* := i^* - 2$ maintaining $\mu(a_{i_1^*+1}) = a_{i_1^*+2}$ and $\mu(a_{i_1^*+3}) = a_{i_1^*+4}$. Since P is odd, we must reach some desired i_ℓ^* after repeating the same argument finitely many times.

[29] To see this, remember that a_{i^*} is assumed to be single at μ. By the regularity of μ, thus, a_{i^*+6} cannot be single at μ.

Next suppose $\nu_2 \rhd_{D_2} \nu_1 \rhd_{D_1} \nu$. For $\nu_2 \not\succ_D \mu$ hold, $a \in D_2$ is necessary and in turn requires either of the following: (1) $a_{i^*+2} \in D_1$ so as to make $D_2 = \{a_{i^*}, a_{i^*+1}\}$ possible, or (2) $i^* \in \{6, 13, 20\}$ and $\mu(a_{i^*-6}) \in D_1$ so as to $D_2 = \{a_{i^*}, a_{i^*-6}\}$ possible. Since a_{i^*+3} is assumed to be matched to a_{i^*+4}, (1) could be the case only if $i^* + 2 \in \{6, 13, 20\}$ and $D_1 \supset \{a_{i^*+2}, a_{i^*-4}\}$. Nonetheless, this cannot be the case for the following reason: When $i^* + 2 \in \{6, 13, 20\}$ (i.e., when $i^* \in \{0, 7, 14\}$), a_{i^*-4} would agree to deviate with a_{i^*+2} only if she is single at ν and hence at μ. By assumptions, however, a_{i^*-4} cannot be single at μ.[30] To check the second possibility, remember that if $i^* \in \{6, 13, 20\}$, $\mu(a_{i^*-6}) = \nu(a_{i^*-6})$ is either a_{i^*-7} or a_{i^*-5}. If it is a_{i^*-7}, she cannot be a member of D_1, since since she is matched to her best possible partner, a_{i^*-6} at ν. If $\mu(a_{i^*-6}) = \nu(a_{i^*-6}) = a_{i^*-5}$, the regularity of μ and the assumption of $\mu(a_{i^*-2}) = a_{i^*+1}$ jointly imply that a_{i^*-4} is matched to her best possible parter, a_{i^*+3}, at μ and hence at ν. Therefore, a_{i^*-4} would not agree to deviate with a_{i^*-5} from ν; i.e., $a_{i^*-5} = \mu(a_{i^*-6}) \in D_1$ is impossible. In conclusion, we have established the robustness of (D, ν) up to depth 2 and completed the proof. ■

E SaRD and Pareto Efficiency

In this appendix, we briefly discuss the efficiency properties of SaRD matchings. To begin, remember that any SaRD matching needs to be regular (Proposition 1). We could thus argue that a SaRD matching always meets a minimal efficiency criterion, in the sense that it leaves no mutually-acceptable pair of singles. Nonetheless, a SaRD matching is not necessarily Pareto efficient, and moreover, the SaRD property is not necessarily preserved by Pareto improvements, as illustrated in the next example:

Example 5. Let $N = \{a_1, a_2, a_3, b_1, b_2, b_3, c_1, c_2, c_3\}$, and let $\succ$ be such that the associated

[30] Since we assume $\mu(a_{i^*-2}) = a_{i^*-1}$ and that all pairs are adjacent at μ, a_{i^*-3} must be single at μ if so is a_{i^*-4}. However, they cannot be simultaneously single by regularity.

party permutation is given by

$$\sigma = \begin{pmatrix} a_1 & a_2 & a_3 & b_1 & b_2 & b_3 & c_1 & c_2 & c_3 \\ a_2 & a_3 & a_1 & b_2 & b_3 & b_1 & c_2 & c_3 & c_1 \end{pmatrix},$$

where the right-hand side denotes $\sigma(a_1) = a_2$, $\sigma(a_2) = a_3$, and so on. It is easy to check such σ induces $\mathscr{P}(\sigma) = \{\{a_1, a_2, a_3\}, \{b_1, b_2, b_3\}, \{c_1, c_2, c_3\}\}$. Suppose further that

- $a_2 \succ_{a_3} b_1 \succ_{a_3} a_3$ and $a_3 \succ_{b_1} b_2$ (i.e., a_3 and b_1 are mutually-acceptable),
- $b_3 \succ_{b_2} c_3 \succ_{b_2} b_1$ and $c_2 \succ_{c_3} b_2 \succ_{c_3} c_3$, (i.e., b_2 and c_3 are mutually-acceptable), and
- all the other pairs of agents across parties (i.e., pairs such as (a_1, b_2), (b_3, c_1), etc.) are not mutually-acceptable.

In this problem, consider two matchings

$$\mu := \{\{a_1, a_2\}, \{a_3\}, \{b_1, b_2\}, \{b_3\}, \{c_1, c_2\}, \{c_3\}\}, \text{ and}$$
$$\mu' := \{\{a_1, a_2\}, \{a_3, b_1\}, \{b_2, c_3\}, \{b_3\}, \{c_1, c_2\}\}.$$

Note that μ is an outcome of Algorithm A and thus is SaRD up to depth 1.[31] The point here is that μ' is *not* SaRD up to depth 1, even though it Pareto-dominates μ. To see this, consider two deviations by $D = \{b_2, b_3\}$, $(\{b_2, b_3\}, \nu)$ from μ and $(\{b_2, b_3\}, \nu')$ from μ', where

$$\nu := \{\{a_1, a_2\}, \{a_3\}\{b_1\}, \{b_2, b_3\}, \{c_1, c_2\}, \{c_3\}\}, \text{ and}$$
$$\nu' := \{\{a_1, a_2\}, \{a_3, b_1\}, \{b_2, b_3\}, \{c_1, c_2\}, \{c_3\}\}.$$

When the original matching is μ, b_1 is left single after $\{b_2, b_3\}$ deviates, and thus, b_1 and b_3 can form a deviation from ν leaving b_2 single; that is, $(\{b_2, b_3\}, \nu)$ is not a robust deviation from μ up to depth 1. In contrast, when $\{b_2, b_3\}$ deviates from μ', b_1 is

[31] More specifically, Algorithm A outputs μ if we choose a_1, b_1, and c_1 as "$a \in P$" in Phase 1, respectively, from $P = \{a_1, a_2, a_3\}, \{b_1, b_2, b_3\}$, and $\{c_1, c_2, c_3\}$.

matched to $c_3 \succ_{b_1} b_3$ and thus would not deviate with b_3. Indeed, the only deviation from ν' is the one by $\{c_2, c_3\}$, which leaves the pair (b_2, b_3) intact. We can thus conclude that $(\{b_2, b_3\}, \nu')$ is a robust deviation from μ' up to depth 1. □

Due to the problems illustrated in the above example, it is difficult to investigate the compatibility of the SaRD property with Pareto efficiency in general. When the problem is simple enough in a certain sense, however, we can guarantee the efficiency of a SaRD matching we construct by Algorithms A–B, and consequently, can establish the following result.

Proposition 10. *Suppose that $(N, \succ)$ is such that for each agent a, the number of acceptable agents to her (i.e., the cardinality of $\{b \in N : b \succ_a a\}$) is no greater than 2. Then, there exists a matching that is SaRD up to depth 3 and Pareto efficient. Further, if $(N, \succ)$ also meets $\#(N, \succ) \leq 5$ (resp. $\#(N, \succ) \leq 3$), there exists a matching that is SaRD up to depth 2 (resp. depth 1) and Pareto efficient.*

Proof. Let σ be a party permutation for $(N, \succ)$ and μ a regular matching satisfying Properties 1–2. Given Propositions 4–7, it suffices to establish the Pareto efficiency of μ. Towards a contradiction, suppose that ν Pareto dominates μ, and hence that there is $a \in I^{\circ}_{\mu}$ such that $\nu(a) \succ_a \mu(a)$ by Lemma 1. By Property 2, $P(a)$ must be an odd party. First, suppose that $P(a)$ is non-solitary. Then, $\nu(a)$ should be either $\pi(a)$ or $\sigma(a)$, since no other agent is acceptable to a under the assumption on $\succ$. By the regularity of μ, $\pi(a)$ and $\sigma(a)$ are matched, respectively, to $\pi^2(a)$ and $\sigma^2(a)$ at μ. Note also that $\pi(a)$ and $\sigma^2(a)$ are the best possible partners, respectively, for $\pi^2(a)$ and $\sigma(a)$. Therefore, $\pi^2(a)$ should prefer μ to ν if $\nu(a) = \pi(a)$, and $\sigma(a)$ should prefer μ to ν if $\nu(a) = \sigma(a)$; however, this is a contradiction to the assumption that ν Pareto dominates μ. Second, suppose that $P(a)$ is solitary. In this case, $|P(\nu(a))| = 2$ is necessary for a to be acceptable for $\nu(a)$, and if so, $\nu(a)$ should prefer $\mu(\nu(a)) = \pi(\nu(a))$ to a, as a should be inferior by the definition of a party permutation.[32] This again contradicts

[32]Remember that when $P(a)$ is solitary and hence $\pi(a) = a$, being acceptable for a is equivalent to being superior for a.

the assumption that ν dominates μ, and we complete the proof. ■

F Weak Stability against Robust Deviations

In this appendix we discuss an alternative, weaker version of our solution concept.[33] Recall that the original definition of robust deviations, requires $\nu_\kappa \succeq_D \mu$ for any sequence $(D_1, \nu_1), \ldots, (D_\kappa, \nu_\kappa)$ of subsequent deviations satisfying $(*)$. Alternatively, one could argue that $a \in D$ would hesitate to form the original deviation (D, ν) when she is indifferent between ν_κ and μ, if there is some (infinitesimally) small cost to form a deviation. To investigate such a scenario, let us call a deviation (D, ν) from μ *strongly robust up to depth k* if it satisfies $\nu_\kappa \succ_D \mu$ for any sequence for any $\kappa \leq k$ and any sequence $(D_1, \nu_1), \ldots, (D_\kappa, \nu_\kappa)$ satisfying $(*)$. Correspondingly, we say a matching μ to be *weakly SaRD up to depth k*, if no deviation from μ is strongly robust up to depth k. By definition, a matching is weakly SaRD up to depth k if it is SaRD up to depth k.

With this weaker requirement, actually, we can always construct a matching that is weakly SaRD up to depth $k = 1$. In doing so, we first provide a sufficient condition for a matching to be weakly SaRD up to depth 1:

Lemma 9. *Suppose that μ is an individually rational matching satisfying the following conditions for all $a \in N$:*

- *if a is in an odd party (i.e., $a \in P \in \mathscr{P}(\sigma)$ and $|P|$ is odd), $\mu(a)$ is inferior for a; and*
- *if a is in an even party (i.e., $a \in P \in \mathscr{P}(\sigma)$ and $|P|$ is even), $\mu(a) \succeq_a \pi(a)$.*

Then, such a matching μ is weakly SaRD up to depth 1.

Proof. Towards a contradiction, suppose that μ is not weakly SaRD up to depth 1; i.e., there is a deviation (D, ν) that is strongly robust up to depth 1. Since μ is assumed to be individually rational, so is ν. Throughout the remainder of the proof, let N_o and N_e be, respectively, the members of odd parties and even parties.

[33] The results in this appendix are originally shown in Kasuya and Tomoeda (2012).

We first show $D \cap N_o \neq \varnothing$. If $a \in D \cap N_e$, then $\nu(a)$ is superior for a, since by assumptions, $\nu(a) \succ_a \mu(a) \succeq_a \pi(a)$. By the definition of a party permutation, a must be inferior for $\nu(a)$. This implies that $\nu(a)$ is a member of N_o, since otherwise she should prefer $\mu(\nu(a)) \in \{\pi(\nu(a)), \sigma(\nu(a))\}$ to $\nu(\nu(a)) \equiv a$. Therefore, $D \subseteq N_e$ is impossible.

Now let $D_S \subseteq D$ (resp. $D_I \subseteq D$) be the set of $a \in D$ such that $\nu(a)$ is superior (resp. inferior) for a. By definition, $D_S \cup D_I = D$ and $D_S \cap D_I = \varnothing$. Note that $(D \cap N_e) \subseteq D_S$ as argued in the previous paragraph, and that $|D_I| \geq |D_S|$ follows from the definition of a party permutation. Therefore, $|D_I \cap N_o| \geq |D_S \cap N_o|$ must hold. Combined with $D \cap N_o \neq \varnothing$, it also follows that $D_I \cap N_o \neq \varnothing$.

Next, take an arbitrary $a \in D_I \cap N_o$. Then a cannot be a member of a solitary party, i.e., $\{a\} \notin \mathscr{P}$.[34] Further, we can check $\sigma(a) \in D_S$ as follows: Note first that $\nu(a) \neq \sigma(a)$ by the assumption of $a \in D_I$. If $\nu(\sigma(a))$ is inferior for $\sigma(a)$, then $a = \pi(\sigma(a)) \succ_{\sigma(a)} \nu(\sigma(a))$ as well as $\sigma(a) \succ_a \nu(a)$. Thus we can take a new matching ν' by matching a and $\sigma(a)$ so that $(\{a, \sigma(a)\}, \nu')$ forms a deviation from ν. It follows from the individual rationality of μ that $\mu(\nu(a)) \succeq_{\nu(a)} \nu(a) = \nu'(\nu(a))$, which contradicts the strong robustness of (D, ν). Therefore, $\nu(\sigma(a))$ must be superior for $\sigma(a)$; that is, $\sigma(a) \in D_S$. Analogously, we can also verify $\pi(a) \in D_S$: Otherwise $\{a, \pi(a)\}$ forms a deviation ν' and leads to a contradiction with the strong robustness of (D, ν).

In the previous paragraph, we have shown that if $a \in D_I \cap N_o$, she is not in a solitary party and $\sigma(a), \pi(a) \in D_S \cap N_o$. Therefore, $|D_I \cap P| \leq |D_S \cap P|$ holds for each odd party $P \in \mathscr{P}(\sigma)$. Since $D_I \cap N_o \neq \varnothing$, further, the strict inequality holds for at least one non-solitary odd party. Summing these inequalities across the odd parties, we obtain $|D_I \cap N_o| < |D_S \cap N_o|$, but this is a contradiction because, as mentioned above, the definition of a party permutation implies $|D_I \cap N_o| \geq |D_S \cap N_o|$. ■

With the sufficient condition above, it is straightforward in any problem to con-

[34] If $\{a\} \in \mathscr{P}$, then $\pi(a) = a$ and hence, $a \in D_I$ is followed by $a \succeq_a \nu(a) \succ_a \mu(a)$. However, this contradicts the individual rationality of μ.

struct a weakly SaRD matching:

Theorem 4. *For any roommate problem $(N, \succ)$, there exists a matching that is weakly SaRD up to depth* 1.

Proof. Fix a problem and a party permutation σ. Construct a matching μ as follows: For each odd party $P \in \mathscr{P}(\sigma)$ and for each $a \in P$, let $\mu(a) = a$. For each even party $P' \in \mathscr{P}(\sigma)$, order its elements as $P' = \{a_1, a_2 \ldots, a_{2m}\}$ so that $\sigma(a_{2j-1}) = a_{2j}$ for each $j \in \{1, \ldots, m\}$ and let $\mu(a_{2j-1}) = a_{2j}$ for each $j \in \{1, \ldots, m\}$. This μ is individually rational and satisfies the conditions in Lemma 9. It is thus weakly SaRD up to depth 1. ■

In the above proof, we leave all odd-party members unmatched so as to apply Lemma 9. This is *not always* necessary and there can exist a weakly SaRD matching up to depth 1 where some odd-party members are matched:

Example 6. Let $N = \{1, 2, 3\}$ and $\succ_i$ be such that $(i+1) \succ_i (i-1) \succ_i i \pmod 3$ for each $i \in N$. Define three matchings μ, ν and ν', respectively, by $\mu = \{\{1, 2\}, \{3\}\}$, $\nu = \{\{1\}, \{2, 3\}\}$ and $\nu' = \{\{1, 3\}, \{2\}\}$. In this problem, μ is weakly SaRD up to depth 1: the only deviation from μ is $(\{2, 3\}, \nu)$, but this is not strongly robust up to depth 1 because $\nu' \rhd_{\{1,3\}} \nu$ and $\mu(2) = 1 \succ_2 2 = \nu'(2)$. Symmetrically, ν and ν' are also weakly SaRD up to depth 1. □

At the same time, however, it is *sometimes* necessary to unmatch all odd-party members as in the next example. Consequently, there may not exist a regular matching that is weakly SaRD up to depth 1.

Example 7. Let $N = \{1, 2, 3, 4, 5\}$ and for each $i \in N$, let $\succ_i$ be such that

- only $i + 1$ and $i - 1 \pmod 5$ are acceptable for i, and
- $(i+1) \succ_i (i-1) \succ_i i \pmod 5$.

In this problem, it is easy to check that the (unique) party permutation σ is given by $\sigma(i) = (i+1) \pmod 5$, and hence by Lemma 9, $\mu = \mathrm{id}$ is weakly SaRD up to

depth 1. Actually, we can check it is the only such matching as follows: Consider two matchings $\mu_1 = \{\{1,2\},\{3\},\{4\},\{5\}\}$ and $\mu_2 = \{\{1,2\},\{3,4\},\{5\}\}$. Note that $(\{4,5\},\nu)$ is a deviation both from μ_1 and from μ_2, where $\nu = \{\{1,2\},\{3\},\{4,5\}\}$. The only deviation from ν is $(\{2,3\},\nu')$ with $\nu' = \{\{1\},\{2,3\},\{4,5\}\}$, and both 4 and 5 are strictly better off at ν' than either at μ_1 or at μ_2. That is, $(\{4,5\},\nu)$ is a strongly robust deviation up to depth 1 either from ν_1 or ν_2, and thus, neither μ_1 nor μ_2 is weakly SaRD up to depth 1. All the other cases are symmetric to either μ_1 or μ_2. □

G Histroty-Dependant Rational-Expectation Farsighted Stable Sets

This appendix considers the *history-dependent rational-expectation farsighted stable set (HREFS)* of Dutta and Vartiainen (2020) in the roommate problem. Specifically, we present a class of examples where all individually rational matchings are qualified to be stable according to a HREFS while not all of them are SaRD.

To define the relevant concepts, first fix an arbitrary $(N,\succ)$ and let X denote the set of all individually rational matchings. Given two distinct $\mu,\nu \in X$, let $\mathscr{E}(\mu,\nu) \subset 2^N$ be the *effectiveness relation* between μ and ν: $D \in \mathscr{E}(\mu,\nu)$ if (i) $a \in D \Rightarrow \mu(a) \neq \nu(a) \in D$, (ii) $[b \notin D \text{ and } \mu(b) \in D] \Rightarrow \nu(b) = b$, and (iii) $c, \mu(c) \notin D \Rightarrow \nu(c) = \mu(c)$.[35] That is, $D \in \mathscr{E}(\mu,\nu)$ denotes the fact that the set D of agents can enforce by themselves the change from μ to ν. A sequence $p = (\mu_0, D_1, \mu_1, \ldots, D_k, \mu_k)$ is an *objection path* if $\mu_0,\ldots,\mu_k \in X$, $D_\kappa \in \mathscr{E}(\mu_{\kappa-1},\mu_\kappa)$ for all $\kappa \in \{1,\ldots,k\}$, and $\mu_k \succ_{D_\kappa} \mu_{\kappa-1}$ for all $\kappa \in \{1,\ldots,k\}$. Sequences of length one, such as $p = (\mu_0)$, are also considered as an objection path. Given an objection path p, let $\iota(p)$ and $\tau(p)$ denote the initial and terminal states of p; i.e., if $p = (\mu_0, D_1, \mu_1, \ldots, D_k, \mu_k)$, then $\iota(p) = \mu_0$ and $\tau(p) = \mu_k$.

Definition 4 (Dutta and Vartiainen, 2020)**.** A set P of objection paths is *coherent* if it satisfies the following:

[35] Note that $D \in E(\mu,\nu)$ differs from $\mu \rhd_D \nu$ in that the former does not require $\mu \succ_D \nu$.

- for every matching μ, there is $p \in P$ such that $\iota(p) = \mu$,
- if $p = (\mu_0, D_1, \mu_1, \ldots, D_k, \mu_k) \in P$, then $p' = (\mu_1, D_2, \ldots, D_k, \mu_k) \in P$,
- if $p = (\mu_0, D_1, \mu_1, \ldots, D_k, \mu_k) \in P$, then for any ν such that $D_1 \in \mathscr{E}(\mu_0, \nu)$, there exists $q \in P$ such that $\iota(q) = \nu$ and $\tau(q) \not\succ_{D_1} \mu_k$, and
- if $(\mu) \in P$, then for any D and ν such that $D \in E(\mu, \nu)$, there exists $q \in P$ such that $\iota(q) = \nu$ and $\tau(q) \not\succ_D \mu$.

When P is a coherent set of objection paths, $\tau(P) := \{\tau(p) : p \in P\}$ is a *history-dependent rational-expectation farsighted stable set* (*HREFS* for short).[36] □

Dutta and Vartiainen (2020) show that a HREFS exists for any finite games and hence, for any roommate problem. However, a HREFS may not be very useful in the roommate problems, for it can be "too inclusive." The following example illustrates this point in the simplest case of a single odd party.

Example 8. Suppose $N = \{a_1, \ldots, a_n\}$ with n being odd and that for each $i \in \{1, \ldots, n\}$, a_i's preference $\succ_{a_i}$ is such that $a_{i+1} \succ_{a_i} a_{i-1}$ and all the others are unacceptable, where the subscripts are in modulo n. Let P be the set of all objection paths, including (μ) for all individually rational μ. In what follows, we show that this P is coherent and thus that the set of all individually rational matchings constitutes a HREFS. As the first two requirements for coherency is straightforward, we only prove the latter two.

To establish the third requirement, let $p = (\mu_0, D_1, \mu_1, \ldots, D_k, \mu_k)$ and $D_1 \in \mathscr{E}(\mu_0, \nu)$. First, suppose that $\{a_i, a_{i+1}\} \subset D_1$ for some i. Since $\mu_k \succ_{D_1} \mu_0$ and μ_0 is individually rational, neither a_i nor a_{i+1} is single at μ_k. By the definition of $\succ$, then, one of a_i and a_{i+1} is matched at μ_k to their best possible partner (respectively, a_{i+1} and a_{i+2}). Therefore, $\tau(q) \succ_{D_1} \mu_k$ never holds for any matching $\tau(q)$. Next, suppose that D_1 contains no adjacent pair. In this case, the agents in D_1 cannot form a new pair themselves; that is, $\nu(a) = a$ for any $a \in D_1$ and any ν such that $D \in \mathscr{E}(\mu, \nu)$. This implies $\mu_1 = \nu$ and hence, $q = (\mu_1, D_2, \ldots, D_k, \mu_k) \in P$ satisfies both $\iota(q) = \nu$ and $\tau(q) \not\succ_{D_1} \mu_k$.

[36] Strictly speaking, this is the characterization, rather than the definition, of a HREFS (Dutta and Vartiainen, 2020, Theorem 1). We omit the original definition for conciseness, as it is easier to establish our point based on the characterization.

To confirm the fourth requirement, let μ be individually rational and $D \in \mathscr{E}(\mu, \nu)$. First, suppose that there is no $a_i \in D$ such that $\nu(a_i) = a_{i+1}$. If so, we must have $\nu(a) = a$ for all $a \in D_1$ and hence, $\mu \succeq_N \nu$. That is, $q = (\nu) \in P$ satisfies both $\iota(q) = \nu$ and $\tau(q) \not\succ_D \mu$. Next, suppose that $\nu(a_i) = a_{i+1}$ for some $a_i \in D$. Let k be the smallest positive integer such that $\nu(a_{i+2k}) \neq a_{i+2k+1}$, which should exist by the oddness of n. Note that a_{i+2k} must be single at ν by definition. Let $D_\kappa = \{a_{i+2(k-\kappa)+1}, a_{i+2(k+1-\kappa)}\}$ for each $\kappa \in \{1, \ldots, k\}$, and construct $\nu_1, \ldots, \nu_k$ by recursively matching D_κ so that $\nu_k \rhd_{D_k} \nu_{k-1} \ldots \nu_1 \rhd_{D_1} \nu$. Note that $q = (\nu, D_1, \nu_1, \ldots, D_k, \nu_k)$ is an objection path. Since $a_i \in D$ is single at ν_k, we have $\tau(q) = \nu_k \not\succ_D \mu$ and the proof is complete. $\square$