Graduate School of Economics Discussion Paper No. 2019-03

Stability against Robust Deviations in the Roommate Problem

Daisuke Hirata[*]
Hitotsubashi University

Yusuke Kasuya[†]
Kobe University

Kentaro Tomoeda[‡]
University of Technology Sydney

May 15, 2019

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

1 Introduction

The *roommate problem* is the one-sided one-to-one matching problem. On the one hand, it is the simplest class of one-sided matching problems, and is a special case of both hedonic coalition formation (Bogomolnaia and Jackson, 2002) and network formation (Jackson, 2008). On the other hand, it is general enough to capture some difficulties associated with one-sidedness: In particular, the roommate problem may not possess a stable matching even though the marriage problem (i.e., the two-sided one-to-one matching problem) is its subclass and always has a stable matching.[1] For the non-existence of a stable outcome is also an issue in the more-general one-sided problems, the roommate problem has been long studied in game theory and other related fields.

The purpose of the present paper is to propose a (class of) new solution concept(s) for the roommate problem, which weakens stability and is applicable even when no stable matching exists. In doing so, 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

[1] Moreover, it is shown by simulations that the proportion of the 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).

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 $v_1^1, \dots, v_1^3$ and $\nu_2^1, \dots, \nu_2^6$ 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 $v_1^1, \dots, v_1^3$ but at least one does at some of $\nu_2^1, \dots, \nu_2^6$. 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 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 one could 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. 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 would be then natural to argue that potential deviators would have less hesitation to realize a deviation in the former case than the latter. 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 make a matching lesser stable than the others, 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 a_1 and a_3 are matched to a_2 and a_4, respectively, and a_5 is single. 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.[2] From these observations, one might expect that it becomes harder to eliminate serious deviations (i.e., those robust up to larger k's) when there is a longer 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.[3] 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 6.1.

[2] See Example 1 in Section 6.1 for a formal proof.

[3] 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 Section 6.2.

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 presents our algorithm to construct SaRD matchings and its key properties. Section 4 demonstrates some implications of those properties, and then Section 5 provides the main results. Section 6 further discusses our concepts and results.

1.1 Related Literature

In the matching and related literatures, we are not the first to define a stability concept based on chains of deviations and their final outcomes, and a number of related studies take a similar approach. 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 interpretation 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 (2009) 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 Section 6.3.

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 details in Section 6.4.

2 Notation and Definitions

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 $\mu = \{\{1, 2\}, \{3\}\}$, it refers to the matching defined by $\mu(1) = 2$, $\mu(2) = 1$, and $\mu(3) = 3$. Given a subset $D \subseteq N$ of agents and two matchings 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 $a, b \in N$. Note that 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 (i) $a \in D \Rightarrow \nu(a) \in D$, (ii) $[b \notin D$ and $\mu(b) \in D] \Rightarrow \nu(b) = b$, (iii) $c, \mu(c) \notin D \Rightarrow \nu(c) = \mu(c)$, and (iv) $\nu \succ_D \mu$. 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 (D, ν) to be a deviation from an individually rational μ with $D = \{a, b\}$, 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 in what follows we do not fully specify the associated ν when $|D| = 2$, it should thus cause no confusion. When (D, ν) is a deviation from μ, we write $\nu \rhd_D \mu$. 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

$$\nu_\kappa \rhd_{D_\kappa} \nu_{\kappa-1} \rhd_{D_{\kappa-1}} \ldots \rhd_{D_2} \nu_1 \rhd_{D_1} \nu. \tag{$*$}$$

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.[4] By definition, if a deviation is robust

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

up to depth k, then so is it up to any $k' < k$. Consequently, if a matching is SaRD up to depth k, then it is also SaRD up to depth k' for any $k' > k$.

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 definitions per se. In response to such a concern, we make two remarks. First, requiring consistency could lead to some conceptual subtlety, making it difficult for our 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 might be unable to determine pointewise 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 one.

Second but not less importantly, we do not claim that a SaRD matching is fully stable in any sense 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 this 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.[5] Further, even if she believes the others to be sophisticated

for stability).

[5] This scenario parallels with the level-k theory, where each agent is assumed to believe the others are of lower levels of strategic sophistication than herself.

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.

2.1 Tan’s (1991) Concepts and Results

In this subsection, we introduce the concepts and results by 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.[6] Throughout the rest of the paper, given a permutation σ over N, let π denote its inverse σ^{-1}.

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:

- $|P| = 1$;
- $|P| = 2$ 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, a is said to be *inferior* for b.[7]

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. □

[6] 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 $\sigma^n(a_1) = a_1$.

[7] 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.

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

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

That is, $\#(N, \succ)$ denotes the maximal size of odd parties in $(N, \succ)$ if there exists any, and is zero otherwise. Note that this definition is independent of the choice of σ thanks to the above theorem. Tan (1991) characterizes the existence of a stable matching as follows:

Theorem (Tan, 1991)**.** *A stable matching exists in a roommate problem $(N, \succ)$ if and only if $\#(N, \succ) \leq 1$.*[8]

[8] For a generalization to weak preferences, see also Chung (2000).

3 The Algorithm

In this section, we introduce our algorithm that computes a matching μ, from an arbitrarily given problem $(N, \succ)$ and an associated party permutation σ as its inputs.[9] While the procedure of the algorithm we provide in Section 3.1 could appear complicated, its goal is simple: it is designed to guarantee that its outcome μ always satisfies five key properties as well as regularity.

Throughout the rest of the paper, we fix an arbitrary party permutation σ and take it as given. Even though many of our concepts and symbols (such as superiority, π, $P(a)$, etc.) implicitly depend on the choice of σ, it should thus cause no confusion. Now, to describe those properties, let us define

$$I_\mu^\circ := \{a \in N : \pi(a) \succ_a \mu(a)\},$$

taking a matching μ as given. Notice that for $a \notin I_\mu^\circ$, $\nu(a) \succ \mu(a)$ implies $\nu(a)$ being superior for a. Put differently, I_μ° is the set of agents who can potentially deviate with inferior agents for them. The following are the five key properties we will need. We will exploit the first four to warrant SaRD up to depth $k = 3$, and the last one to establish sufficient conditions for $k = 1$ and 2.[10]

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

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

Property 3. *For any $a \in I_\mu^\circ$, $\sigma^2(a) \notin I_\mu^\circ$.*

Property 4. *For any $a \in I_\mu^\circ$, $\mu(\sigma(a)) = \sigma^2(a)$ and $\mu\left(\sigma^3(a)\right) = \sigma^4(a)$ imply [1] $\mu\left(\sigma^5(a)\right) = \sigma^6(a)$ if $|P(a)| = 7$, and [2] $\sigma^5(a) \in I_\mu^\circ \not\ni \sigma^6(a)$ if $|P(a)| > 7$.*

[9] Although we take the party permutation σ as given, Tan and Hsueh (1995) provide an algorithm to compute a party permutation in $O(|N|^2)$-time.

[10] Indeed, the complexity of the algorithm is mostly due to Property 4. If we give it up, we can construct a simpler algorithm, whose outcome satisfies the other Properties and is always SaRD up to depth $k = 4$.

Property 5. *For any $a \in I_\mu^\circ$,*

$$\Big[|P(a)| = 3 \text{ or } \Big\{|P(a)| = 5 \text{ and } \mu(\sigma(a)) = \sigma^2(a)\Big\}\Big] \Longrightarrow \mu(\pi(a)) = \pi^2(a).$$

3.1 Description of the Algorithm

Taking a problem $(N, \succ)$ and a party permutation σ as given, construct a matching μ as follows.[11] 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.

3.1.1 Phase 1

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 $j \in \left\{1, \dots, \frac{|P|}{2}\right\}$, as illustrated in Figure 2 (a).

3.1.2 Phase 2

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 $a \in E$ and define $\mu\left(\sigma^{3j}(a)\right) = \sigma^{3j+1}(a)$ for each $j \in \left\{1, \dots, \frac{|P|}{3}\right\}$, 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 $P(a) \in \mathscr{O}_{3\times}$*

[11] Note that the outcome of our algorithm below is not uniquely pinned down, as it can vary depending on how to take $a \in P$ in Phases 1, 2, and 4, and how to order the members of U_0 and R_0 in Phases 3 and 5. However, our main results apply to any of those possible outcomes. See also Section 6.1 for further discussions.

and a is not matched in this phase, then $\pi(a)$ and $\sigma(a)$ are matched, respectively, to $\pi^2(a)$ and $\sigma^2(a)$. □

3.1.3 Phase 3

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.[12] Arbitrarily order the members of U_0 as $x_1, \dots, x_{|U_0|}$ and iterate the following step for $t = 1, \dots, |U_0|$.

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 $t = 1, \dots, |U_0|$ 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 $\mathscr{U}_t = \mathscr{U}_{t-1}$.[13] 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 $q, r \in \{1, \dots, |P(x_t)|\}$ 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

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

[13] 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.

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 $j \in \{1, \ldots, m-1\}$, and $\mu((\sigma^{2m}(y_t)) = \sigma^{2m+1}(y_t)$, leaving $\mu\left(\sigma^{2m-1}(y_t)\right)$ undefined. Figure 3 illustrates 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 4 illustrates 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 5 (a).

- Matching among $P(x_t) \in \mathscr{U}_{t-1}$:[14]

  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 $\mu\left(\sigma^{3n}(x_t)\right) =$

[14] 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.

$\sigma^{3n+1}(x_t)$. Figures 5 (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 $\mu\left(\pi(x_{t'})\right) \neq \pi^2(x_{t'})$ only if $x_t = \pi(x_{t'})$ is matched to some y_t an earlier step $t < t'$ such that $P(x_t) \in \mathscr{U}_{t-1}$.*[15] □

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 $x_{t'} \in \left\{\pi(x_t), \pi^2(x_{t'})\right\}$ and thus $P(x_t) = P(x_{t'}) \in \mathscr{U}_{|U_0|}$.* □

3.1.4 Phase 4

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 6:

- 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 $j \in \{3, \ldots, n\}$.[16]
- 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 $j \in \{2, \ldots, n\}$.

[15] For instance, take $x_{t'}$ to be w_1 in Figure 5 (b).

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

Remark 5. *As illustrated in Figure 6, if $P(a) \in \mathscr{V}$ but a is not matched in this phase, $\pi(a)$ and $\sigma(a)$ are matched, respectively, to $\pi^2(a)$ and $\sigma^2(a)$. Combined with Remarks 1 and 3, if $a \in I_\mu^\circ$ at the final outcome,*

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

Remark 6. *If a and b 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, $a \in \{\pi(b), \pi^2(b)\}$ or $b \in \{\pi(a), \pi^2(a)\}$, as argued in Remark 4, but not both of such a and b can remain unmatched after Phase 4 matches the agents in $P(a) = P(b)$ as specified above.* □

3.1.5 Phase 5

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

Step $\boldsymbol{\tau = 1, \dots, |R_0|}$ 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 $R_\tau := R_\tau - \{r_\tau, r_i\}$.[17] 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 $\mu(r) = r$.

[17] 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.

3.2 Key Properties of the Algorithm

As mentioned at the beginning of this Section, the above algorithm is designed so that its outcomes always satisfy Properties 1–5. Here we formally establish this fact.

Proposition 1. *Let μ be an outcome of the algorithm of Section 3.1. Then, it is regular and satisfies Properties 1–5 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 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 $b \in \Sigma_t$.[18] 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

[18] 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'.

$\mu(a) \in \{\pi(a), \sigma(a)\}$ 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 $a \in I^{\circ}_{\mu}$, 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. ■

Proof of Property 4. 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 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 5 (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 6 (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 6 (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 5. Suppose $a \in I^{\circ}_{\mu}$, where μ is an outcome of the algorithm. First, if $|P(a)| = 3$, Phase 2 should match $\pi(a)$ and $\pi^2(a)$. Second, suppose that $|P(a)| = 5$ and $\mu(\pi(a)) \neq \pi^2(a)$. As argued in Remark 5, then, $\pi(a) = x_t$ is matched to y_t in some step t of Phase 3 with $P(a) \in \mathscr{U}_t$. More specifically, $y_t = \sigma(a)$ is the only possibility under the assumption of $|P(a)| = 5$.[19] It then follows that $\mu(\sigma(a)) = \pi(a) \neq \sigma^2(a)$ as required. ■

3.3 Performance of the Algorithm when a Stable Matching Exists

Although our goal is to establish a global property that is applicable even when no stable matching exists, it should be noted that the outcomes of our algorithm are stable whenever a stable matching exists.

Proposition 2. *For any roommate problem $(N, \succ)$ with a stable matching, any outcome μ of the above algorithm is stable.*

Proof. Suppose that a stable matching exists in $(N, \succ)$ and σ is a party permutation. By Tan's (1991) Theorems we stated in Section 2.1, thus, each party $P \in \mathscr{P}(\sigma)$ is either even or solitary. In such a case, only Phase 1 matches agents throughout the entire algorithm. Specifically, any outcome of our algorithm is such that (i) if a is in an even party, $\mu(a) \in \{\pi(a), \sigma(a)\}$, and (ii) if a is in a solitary party, $\mu(a) = a = \pi(a)$.[20] By the definition of a party permutation, such a matching is stable. ■

4 Implications of Properties 1–2

In this section, we provide a number of preliminary lemmas that follow from Properties 1–2. They impose restrictions on possible deviations from a matching satisfying those properties, and as such, will be useful when we establish our main results in the next section. To concisely state those restrictions, here we introduce some more notation. Taking

[19] To see this, note first that if $P(x_t) = 3n+2$ and $y_t \notin P(x_t)$, as illustrated in Figure 5 (c), $\sigma(x_t) \equiv a$ should be matched to $\sigma^2(x_t) \equiv \sigma(a)$. In the case of $P(a) = 5$, thus, $y_t \in P(a)$ is necessary for $a \in I^{\circ}_{\mu}$. Moreover, since neither $\pi(x_t) \equiv \pi^2(a)$ nor $\pi^2(x_t) \equiv \pi^3(a)$ can be a member of Σ_t by definition, $\pi^3(x_t) \equiv \sigma(a)$ is the only candidate for $y_t \in P(a)$.

[20] Remember that when a is a solitary party member (i.e., when $a = \pi(a)$), b is acceptable for a if and only if b is superior for a. By the definition of a party permutation, therefore, no pair of solitary party members is mutually acceptable.

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. Further, divide $D \cap S_\nu$ into two as follows:

$$Cy := \{a \in D \cap S_\nu : (\pi \circ \nu)^t(a) \in S_\nu \text{ for all } t \in \mathbb{N}\}, \text{ and} \tag{1}$$

$$Ch := (D \cap S_\nu) - Cy. \tag{2}$$

Note that by the finiteness of N,

$$[a \in Cy] \implies [\text{there exists } t^* \in \mathbb{N} \text{ such that } (\pi \circ \nu)^{t^*}(a) = a], \tag{3}$$

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, as illustrated in Figure 7. 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.

4.1 Implications of Property 1

The first Lemma is a key implication of Property 1. It guarantees that for any deviation (D, ν), there exists some agent $a \in D \cap I_\mu^\circ$. 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) $a \in I_\mu^\circ$, b is an inferior agent for a, and $\mu(b) \neq b$; and (ii) $b \in I_\mu^\circ$, 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.[21] Third, $\nu \succ_E \mu$ and Property 1 imply *both* [1] either a is inferior for b or

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

$\mu(b) \neq b$, 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 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. ■

4.2 Implications of Property 2

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 $a \in Cy$, then $(\pi \circ \nu)^t(a) \in Cy$ for all $t \in \mathbb{N}$.*

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

Lemma 5. *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 6. *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 $\nu(a) = \mu(a) \in I_\mu^\circ$ 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 4 entails $a \in Cy \subseteq D$, which contradicts the assumption of $a \notin D$. ■

Lemma 7. *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 4. As $b \in D \cap S_\nu$ implies $\nu(b) \in D - S_\nu$, we should have $\nu(b) \in I_\mu^\circ$ 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. ■

5 Main Results

In this section, we establish a bound for the robustness of a deviation ν from μ in each of three mutually-exclusive subcases depending on the composition of $D \cap S_\nu \equiv Ch \cup Cy$ (Claim 1–3). Combining those Claims altogether, we obtain our main results.

Claim 1. Suppose $\nu \rhd_D \mu$, where μ is a regular matching satisfying Property 2. If $D \cap S_\nu = \emptyset$, then ν is not robust up to depth 1.

Proof. 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. ■

Claim 2. Suppose $\nu \rhd_D \mu$, where μ is a regular matching satisfying Properties 1 and 2. If $Ch \neq \emptyset$, then ν is not robust up to depth 1.

Proof. By Lemma 5, 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. ■

Claim 3. Suppose $\nu \rhd_D \mu$, where μ is a regular matching satisfying Properties 1–4. If $Ch = \emptyset \neq Cy$, then ν is not robust up to depth (at most) 3.

Proof. To begin, 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 without loss of generality for the following reason: Such b would not exist only if for each $b \in \nu(Cy)$, there exists some $t_b \in \mathbb{N}$ such that $\sigma^{4t_b}(b) = b$.[22] This

[22] Notice that $\sigma^3(b) \in Cy$ is equivalent to $\sigma^4(b) \in \nu(Cy)$, for $\nu(Cy) = \sigma(Cy)$ by Lemma 4.

cannot be the case, however, since $b \in \nu(Cy)$ implies $\pi(b) \in Cy$ by Lemma 4, and hence, $P(b)$ must be an odd party by Lemma 7. Let $m \in \mathbb{N}$ be such that $|P(b)| = 2m + 1$, and 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.[23] 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 4, 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 3 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 6 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 3 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

[23] Notice that if μ satisfies Property 5, $c_1 \notin D$ and $\nu(c_1) \neq c_2$ together imply $|P(b)| \geq 5$. This is because if $|P(b)| = 3$, Property 5 implies $\mu(c_1) = c_2$, which contradicts $c_1 \notin D$ and $\nu(c_1) \neq c_2$.

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 $c_3 \notin D \cap S_\nu \equiv Cy \cup Ch$.[24] First, suppose $\nu(c_3) = \mu(c_3) \neq c_4$. Then, as in the last part of Case 1, Lemma 6 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

$$\nu_3 \rhd_{\{b,c_1\}} \nu_2 \rhd_{\{c_2,c_3\}} \nu_1 \rhd_{\{\pi(\nu(c_3)),\nu(c_3)\}} \nu. \tag{4}$$

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 4 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 $c_5 \in I_\mu^\circ$ by Property 4, $c_5 \in S_\nu$ 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 4 implies $c_6 \notin I_\mu^\circ$ 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 $\{c_4, c_5\}$, $\{c_2, c_3\}$,

[24] Note that if μ satisfies Property 5, $c_3 \in S_\nu - D$ implies $|P(b)| > 5$ for the following reason: If $|P(b)| = 5$, Property 5 implies that $c_3 = \pi^2(b)$ and $c_4 = \pi(b)$ are matched with each other at μ. However, this is a contradiction because $c_4 = \pi(b)$ is a member of $Cy \subseteq D$ by definition.

and $\{b, c_1\}$, so that

$$\nu_3 \rhd_{\{b,c_1\}} \nu_2 \rhd_{\{c_2,c_3\}} \nu_1 \rhd_{\{c_4,c_5\}} \nu. \tag{5}$$

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

Combining the three Claims above, we obtain our main theorem:

Theorem 1. *A regular matching satisfying Properties 1–4 is SaRD up to depth (at most)* 3. *For any problem* $(N, \succ)$*, thus, there exists a matching that is SaRD up to depth (at most)* 3.

Proof. The statement immediately follows from Claims 1–3 and Proposition 1. ■

Further, the proof of Claim 3 also establishes sufficient conditions for the outcomes of our algorithm to be SaRD up to depth 1 and 2:

Theorem 2. *If* $\#(N, \succ) \leq 3$*, a regular matching satisfying Properties 1–3 and 5 is SaRD up to depth (at most)* 1*. Thus there exists a matching that is SaRD up to depth* 1 *for any problem such that* $\#(N, \succ) = 3$.

Proof. In the proof of Claim 3, we establish without employing Property 4 that ν is not robust up to depth 1 in the case of $c_1 \notin S_\nu$. With Property 5, $c_1 \in S_\nu$ arises only if $\#(N, \succ) > 3$: For the case of $\nu(c_1) \neq c_2$, see footnote 23; for the case of $\nu(c_1) = c_2$, we established $P(b) > 3$ in the main body of the proof. ■

Theorem 3. *If* $\#(N, \succ) \leq 5$*, a regular matching satisfying Properties 1–3 and 5 is SaRD up to depth (at most)* 2*. Thus there exists a matching that is SaRD up to depth* 2 *for any problem such that* $\#(N, \succ) = 5$.

Proof. In the proof of Claim 3, we establish without employing Property 4 that ν is not robust up to depth 2, except for the case where $c_1 \in S_\nu$, $\nu(c_1) = c_2$, and $c_3 \in S_\nu - D$. With Property 5, such a case arises only if $\#(N, \succ) > 5$; see footnote 24. ■

6 Discussions

6.1 Tightness of the Results

In this section we discuss the tightness of our main results. It should be first noted that the bound of $k = 3$ we establish for general existence in Theorem 1 is tight in the sense that no matching is SaRD up to depth 2 for some problems. The next example, which we have informally discussed in the introduction, illustrates this point.

Example 1. Suppose that $N = \{a_1, a_2, \ldots, a_7\}$ and that for each $i = 1, \ldots, 7$, only a_{i+1} and a_{i-1} are acceptable a_i and $a_{i+1} \succ_{a_i} a_{i-1}$, where all subscripts are in modulo 7. The unique party permutation is given by $\sigma(a_i) = a_{i+1}$ for each i so that $\mathscr{P}(\sigma) = \{N\}$. In this problem, no matching is SaRD up to depth 2. To see this, first notice that for a matching to be SaRD up to any depth, it should match three pairs of adjacent agents; if fewer, such a matching is not regular because there must exist a pair of adjacent (i.e., mutually acceptable) singles. It thus suffices to confirm that $\mu = \{\{a_1, a_2\}, \{a_3, a_4\}, \{a_5, a_6\}, \{a_7\}\}$ is not SaRD up to depth 2, as all the other candidates are symmetric. Starting from μ, a deviation chain $\nu_2 \rhd_{D_2} \nu_1 \rhd_{D_1} \nu \rhd_D \mu$ is possible only with $D = \{a_6, a_7\}$, $D_1 = \{a_4, a_5\}$, and $D_2 = \{a_2, a_3\}$. As $D = \{a_6, a_7\}$ prefer $\nu_2 = \{\{a_1\}, \{a_2, a_3\}, \{a_4, a_5\}, \{a_6, a_7\}\}$ to μ, the deviation ν from μ is robust up to depth 2 and hence μ is not SaRD up to depth 2. $\square$

In contrast, our sufficient conditions in Theorems 2–3 are not tight in two different ways: Note that Theorems 2–3 establish the conditions for *any* outcome of our algorithm to be SaRD up to depth $k = 1$ and 2, while our algorithm generally has multiple possible outcomes as mentioned in footnote 11. First, therefore, *some* outcomes may be SaRD up to a smaller k than those guaranteed by Theorems 2–3, even when *not all* possible outcomes are. Second, there may be a matching that is SaRD up to a smaller k than any outcome of our algorithm is. The following examples demonstrate that such possibilities do indeed exist.

Example 2. Suppose that $N = \{a_1, \ldots, a_7, b_1, \ldots, b_3, c_1, \ldots, c_3\}$, and let

$$\sigma = \begin{pmatrix} a_1 & a_2 & \cdots & a_6 & a_7 & b_1 & b_2 & b_3 & c_1 & c_2 & c_3 \\ a_2 & a_3 & \cdots & a_7 & a_1 & b_2 & b_3 & b_1 & c_2 & c_3 & c_1 \end{pmatrix},$$

where the right-hand side denotes $\sigma(a_1) = a_2$, $\sigma(a_2) = a_3$, and so on. Note that $\mathscr{P}(\sigma) = \Big\{\{a_1, \ldots, a_7\}, \{b_1, \ldots, b_3\}, \{c_1, \ldots, c_3\}\Big\}$. Let $\succ$ be a preference profile such that

- $\succ_{a_1}$ is such that $a_2 \succ_{a_1} a_7 \succ_{a_1} b_1 \succ_{a_1} a_1$,
- $\succ_{a_5}$ is such that $c_1 \succ_{a_5} a_6 \succ_{a_5} a_4 \succ_{a_5} a_5$,
- $\succ_{b_1}$ is such that $a_1 \succ_{b_1} b_2 \succ_{b_1} b_3 \succ_{b_1} b_1$,
- $\succ_{c_1}$ is such that $c_2 \succ_{c_1} c_3 \succ_{c_1} a_5 \succ_{c_1} c_1$, and
- for any $d \notin \{a_1, a_5, b_1, c_1\}$, $\succ_d$ is such that $\sigma(d) \succ_d \pi(d) \succ_d d$.

Further, for any $\alpha, \beta \in N$, assume that α is unacceptable for β unless otherwise specified above. In this problem $(N, \succ)$, where σ is the unique party permutation, we compare two outcomes of our algorithm.

To begin, suppose that b_2 and c_2 are taken as $a \in P$ during Phase 2, so that they are matched to b_3 and c_3. At the beginning of Phase 3, then, $U_0 = \{a_1, \ldots, a_7, b_1, c_1\}$. First, suppose further that a_1 is taken as x_1. Then, $x_1 = a_1$ is matched to $y_1 = b_1$ at step 1 of Phase 3 and the final outcome of the algorithm will be

$$\mu = \Big\{\{a_1, b_1\}, \{a_2\}, \{a_3, a_4\}, \{a_5, c_1\}\{a_6, a_7\}, \{b_2, b_3\}, \{c_2, c_3\}\Big\},$$

as illustrated in Figure 8 (a). Second, suppose instead that c_1 is taken as x_1. Then, $x_1 = c_1$

is matched to $y_1 = a_5$ at step 1 of Phase 3 and the final outcome of the algorithm will be

$$\mu' = \Big\{\{a_1, a_2\}, \{a_3, a_4\}, \{a_5, c_1\}\{a_6, a_7\}, \{b_1\}, \{b_2, b_3\}, \{c_2, c_3\}\Big\},$$

as illustrated in Figure 8 (b).

Comparing these two matchings, we can observe that not all outcomes of our algorithm are SaRD up to the same depth. On the one hand, μ is not SaRD up to 2, because the deviation by $D = \{a_1, a_2\}$ is robust up to depth 2. On the other hand, μ' is SaRD up to depth 1: There are only two deviations from μ', one by $D = \{b_1, b_3\}$ and the other by $D = \{c_1, c_3\}$, and it is immediate to confirm that neither is robust up to depth 1. □

Example 3. Suppose that $N = \{a_1, \ldots, a_5, b_1, \ldots, b_5\}$ and that for each $i = 1, \ldots, 5$,

- only a_{i+1}, a_{i-1}, b_i are acceptable for a_i and $a_{i+1} \succ_{a_i} a_{i-1} \succ_{a_i} b_i$, and
- only b_{i+1}, b_{i-1}, a_i are acceptable for b_i and $b_{i+1} \succ_{b_i} b_{i-1} \succ_{b_i} a_i$,

where all the subscripts are in modulo 5. The unique party permutation is then given by $\sigma(a_i) = a_{i+1}$ and $\sigma(b_i) = b_{i+1}$ so that $\mathscr{P}(\sigma) = \{\{a_1, \ldots, a_5\}, \{b_1, \ldots, b_5\}\}$. Note that in this problem, any outcome of our algorithm is symmetric to

$$\mu := \Big\{\{a_1, a_2\}, \{a_3, a_4\}, \{b_1, b_2\}, \{b_3, b_4\}, \{a_5, b_5\}\Big\},$$

which is graphically illustrated in Figure 9 (a).[25] It is then easy to check that the deviation from μ by $D = \{a_4, a_5\}$ is robust up to depth 1. That is, any outcomes, including μ, of the algorithm are not SaRD up depth $k = 1$, although they are so up to depth $k = 2$ by Theorem 3.

However, the following matching μ', which is illustrated in Figure 9 (b), is SaRD up to

[25] To see this, notice that Phase 3 does not match any agents in this problems, as Σ_t is always empty.

depth 1:

$$\mu' := \Big\{\{a_1, b_1\}, \{a_2, b_2\}, \{a_3, b_3\}, \{a_4, b_4\}, \{a_5, b_5\}\Big\}.$$

Note that for any deviation (D', ν') from μ', there exists either [i] $a_i \notin D'$ such that $\nu'(a_{i-1}) = a_{i-2}$ or [ii] $b_i \notin D'$ such that $\nu'(b_{i-1}) = b_{i-2}$. In the first case, a further deviation ν'_1 by $\{a_i, a_{i-1}\}$ makes $a_{i-2} \in D'$ worse off than at the original μ. Therefore, the deviation (D', ν') is not robust up to depth 1. As the second case is symmetric, we can conclude that μ' is SaRD up to depth 1. □

However, it should be also noted that Theorems 2–3 provide sufficient conditions that depend only on σ, whereas more detailed information of $\succ$ would be necessary to pin down matchings that are SaRD up to the smallest k.[26] Indeed, the next proposition suggests that the sufficient conditions derived from our algorithm are tight among those depend only on σ.

Proposition 3. *Let σ be a permutation over N such that $|P| = 2m + 1$ for some $P \in \mathscr{P}(\sigma)$ and $m \in \mathbb{N}$. Then, there exist $\succ = (\succ_i)_{i \in N}$ and $k \in \{1, 2, 3\}$ such that*

- *σ is a party permutation for $(N, \succ)$,*
- *any outcome of the algorithm in Section 3 for $(N, \succ)$ is SaRD up to depth k, and*
- *no matching is SaRD up to depth $k - 1$ in $(N, \succ)$,*

where "SaRD up to depth 0" should be read as the standard stability.

Proof. Given N and σ, let $\succ$ be such that for all $a, b \in N$, $b \succ_a a \Rightarrow b \in \{\pi(a), \sigma(a)\}$ and $\sigma(a) \succ_a \pi(a)$. It is then immediate to check that σ is a party permutation for $(N, \succ)$. Note also that with such $(N, \succ)$, our algorithm matches no pair of agents during Phases 3 and 5. We consider four possible cases separately.

[26] Notice that a same σ can be a party permutation for various distinct preference profiles.

First, if $\#(N,\succ) = 3$, any outcome of the algorithm is SaRD up to depth 1 by Theorem 2, and no matching is stable by Tan's (1991) Theorem. Thus, the claim holds in this case.

Second, suppose $\#(N,\succ) = 5$ and let $P \in \mathscr{P}(\sigma)$ be a party such that $|P| = 5$. At any regular matching μ, then, the members of P should be matched so that $\mu(a) = \sigma(a)$, $\mu\left(\sigma^2(a)\right) = \sigma^3(a)$, and $\mu\left(\sigma^4(a)\right) = \sigma^4(a)$ for some $a \in P$, as illustrated in Figure 6 (c). It is easy to see that the deviation from μ by $D = \{\sigma^3(a), \sigma^4(a)\}$ is robust up to depth 1. Consequently, no matching is SaRD up to depth 1 while the outcomes of our algorithm are SaRD up to depth 2 by Theorem 3.

Third, suppose $\#(N,\succ) > 5$ and that for any $P \in \mathscr{P}(\sigma)$, $|P|$ is either one, even, or a multiple of three. Fix an arbitrary odd party P with $|P| = 3n$ for some $n > 1$. At any regular matching μ, then, there should exist $\alpha \in P$ (as b_2 in Figure 2 (b) for instance) such that $\mu(\alpha) = \alpha$, $\mu(\pi(\alpha)) = \pi^2(\alpha)$, and $\mu\left(\sigma(\alpha)\right) = \sigma^2(\alpha)$.[27] As the deviation from μ by $D = \{\pi(\alpha), \alpha\}$ is robust up to depth 1, no matching can be SaRD up to depth 1. Now, let μ be an outcome of our algorithm and (ν, D) be an arbitrary deviation from μ such that $D \cap P \neq \varnothing$. Note that for any $\alpha \in D \cap P$, $\mu(\alpha) = \alpha$ and $\nu(\alpha) = \pi(\alpha)$. Moreover, since our algorithm matches the members of P in the way as of Figure 2 (b), either [i] $\nu\left(\sigma(\alpha)\right) = \sigma(\alpha)$ or [ii] $\nu\left(\sigma(\alpha)\right) = \mu\left(\sigma(\alpha)\right) = \sigma^2(\alpha)$ and $\nu\left(\sigma^3(\alpha)\right) = \mu\left(\sigma^3(\alpha)\right) = \sigma^3(\alpha)$. In either case, ν is not robust up to depth 2 and thus, the outcome μ of our algorithm is SaRD up to depth 2.

Lastly, suppose $\#(N,\succ) > 5$ and that $|P|$ is odd but not a multiple of three for some $P \in \mathscr{P}(\sigma)$. Then, at any regular matching μ, there should exist $\alpha \in P$ (as b_4 in Figure 6 (a)–(b) and b_{10} in Figure 6 (d)) such that $\mu(\alpha) = \alpha$, $\mu(\pi(\alpha)) = \pi^2(\alpha)$, $\mu\left(\sigma(\alpha)\right) = \sigma^2(\alpha)$, and $\mu\left(\sigma^3(\alpha)\right) = \sigma^4(\alpha)$. Given this observation, it is easy to check that the deviation ν from μ by $D = \{\alpha, \pi(\alpha)\}$ is robust up to depth 2. That is, no matching is SaRD up to

[27] To see this, remember that for each a, only $\pi(a)$ and $\sigma(a)$ are acceptable. As $|P|$ is odd, there needs to exist some α with $\mu(\alpha) = \alpha$. By regularity, both $\pi(\alpha)$ and $\sigma(\alpha)$ need to be matched, and $\pi^2(\alpha)$ and $\sigma^2(\alpha)$, respectively, are their only possible partners.

depth 2 in this case, whereas the outcomes of our algorithm are always SaRD up to depth 3 by Theorem 1. ■

6.2 SaRD and Efficiency

In this section we briefly discuss the efficiency properties of SaRD matchings. To begin, remember that the outcome of our algorithm is regular (Proposition 1). Indeed, as we formally state below, regularity is a property of a SaRD matching in general, not only of the outcomes of our algorithm. 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.

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

Proof. 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 $c \in N - \{a, b\}$. ■

However, no mutually-acceptable pair of singles is obviously weaker than Pareto efficiency. Then it would be natural to ask if the SaRD property implies full Pareto efficiency, or if the outcomes of the algorithm in Section 3 are Pareto efficient. The answers to these questions are negative: The next example demonstrates that the outcomes of our algorithm are not always Pareto efficient. This is essentially because Pareto improvements do not necessarily preserve the SaRD property (with a same depth k).

Example 4. Let $N = \{a_1, a_2, \ldots, a_7, b_1, \ldots, b_3, c_1, \ldots, c_3, d_1, \ldots, d_3, e_1, \ldots, e_3\}$, and let

$$\sigma = \begin{pmatrix} a_1 & a_2 & \cdots & a_6 & a_7 & b_1 & b_2 & b_3 & \cdots & e_1 & e_2 & e_3 \\ a_2 & a_3 & \cdots & a_7 & a_1 & b_2 & b_3 & b_1 & \cdots & e_2 & e_3 & e_1 \end{pmatrix},$$

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 $\mathscr{P}(\sigma) = \{\{a_1, \ldots, a_7\}, \{b_1, b_2, b_3\}, \ldots, \{e_1, e_2, e_3\}\}$. Let $\succ$ be a preference profile such that

- $\succ_{a_1}$ is such that $a_2 \succ_{a_1} a_7 \succ_{a_1} b_1 \succ_{a_1} a_1$,
- $\succ_{a_5}$ is such that $c_1 \succ_{a_5} a_6 \succ_{a_5} a_4 \succ_{a_5} a_5$,
- $\succ_{b_1}$ is such that $a_1 \succ_{b_1} b_2 \succ_{b_1} b_3 \succ_{b_1} b_1$,
- $\succ_{c_1}$ is such that $c_2 \succ_{c_1} c_3 \succ_{c_1} a_5 \succ_{c_1} c_1$,
- $\succ_{c_2}$ is such that $d_3 \succ_{c_2} c_3 \succ_{c_2} c_1 \succ_{c_2} c_2$,
- $\succ_{c_3}$ is such that $e_1 \succ_{c_3} c_1 \succ_{c_3} c_2 \succ_{c_3} c_3$,
- $\succ_{d_3}$ is such that $d_1 \succ_{d_3} d_2 \succ_{d_3} c_2 \succ_{d_3} d_3$,
- $\succ_{e_1}$ is such that $e_2 \succ_{e_1} e_3 \succ_{e_1} c_3 \succ_{e_1} e_1$,
- for any other agent f, $\succ_f$ is such that $\sigma(f) \succ_f \pi(f) \succ_f f$.

Further, for any $\alpha, \beta \in N$, assume that α is unacceptable for β unless otherwise specified above. In this problem $(N, \succ)$, where σ is the unique party permutation, we compare two matchings illustrated in Figure 10:

$$\mu = \Big\{\{a_1, b_1\}, \{a_2\}\{a_3, a_4\}, \{a_5, c_1\}, \{a_6, a_7\},$$
$$\{b_2, b_3\}, \{c_2, c_3\}, \{d_1, d_2\}, \{d_3\}, \{e_1\}, \{e_2, e_3\}\Big\}, \text{ and}$$
$$\mu' = \Big\{\{a_1, b_1\}, \{a_2\}\{a_3, a_4\}, \{a_5, c_1\}, \{a_6, a_7\},$$
$$\{b_2, b_3\}, \{c_2, d_3\}, \{c_3, e_1\}, \{d_1, d_2\}, \{e_2, e_3\}\Big\}.$$

To begin, note that μ is an outcome of our algorithm, and hence is SaRD up to depth 3 by Theorem 1.[28] Next observe that μ' differs from μ only in that c_2 and c_3 (who are matched with each other at μ) are matched to d_3 and e_1 (who are single at μ). It is then easy to confirm that μ' Pareto-dominates μ, which is SaRD up to depth 3.

However, μ' is not robust up to depth 3. To see this, consider the deviation (D, ν') from μ' by $D = \{a_1, a_2\}$, and suppose that $\nu'_\kappa \rhd_{D_\kappa} \nu'_{\kappa-1} \ldots \rhd_{D_1} \nu$ and $\nu'_\kappa \not\gtrsim_D \mu'$. Notice that a_1 is matched to her best possible partner a_2 at ν, and hence, she does not have an incentive for another deviation unless a_2 is gone. As a_2 prefers only a_3 to a_1, then, we must have $D_\kappa = \{a_2, a_3\}$. Since $\{a_2, a_3\}$ cannot form a deviation directly from μ', it follows that (D, ν') is robust up to depth 1. Following similar arguments, we can further confirm that $D_{\kappa-1} = \{a_4, a_5\}$ and $D_{\kappa-2} \ni c_1$ are also necessary. As neither c_2 nor c_3 would deviate from ν' with c_1, we should have $\kappa > 3$ and hence (D, ν) is robust up to depth 3. □

However, we can guarantee that the outcomes of our algorithm are Pareto efficient when the problem is sufficiently simple in the following sense.

Proposition 4. *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, any outcome of the algorithm in Section 3 is Pareto efficient.*

Proof. Let σ be a party permutation for $(N, \succ)$ and μ an outcome of the algorithm. Under the assumption on $\succ$, only $\pi(a)$ and $\sigma(a)$ are acceptable for a if $|P(a)| > 2$. As a consequence, none of Σ_t is acceptable to x_t in any step t of Phase 3, and no pairs among R_0 are mutually acceptable in Phase 5; that is, no pairs are matched during these two Phases.

Now 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. If $P(a)$ is non-solitary, $\nu(a)$ should be either $\pi(a)$ or $\sigma(a)$, since no other agent is acceptable to a as

[28] For instance, the algorithm outputs μ if one takes b_2, c_2, d_1, and e_2 to be "$a \in P$" in Phase 2, and label $x_1 = a_1$ and $x_2 = c_1$ at the beginning of Phase 3.

mentioned above. However, note that $\sigma(a)$ and $\pi^2(a)$ are matched at μ to their best possible partners, $\sigma^2(a)$ and $\pi(a)$. Therefore, $\sigma(a)$ should prefer μ to ν if $\nu(a) = \sigma(a)$, and $\pi^2(a)$ should prefer μ to ν if $\nu(a) = \pi(a)$. This is a contradiction. If $P(a)$ is solitary, $|P(\nu(a))| = 2$ is necessary for a to be acceptable for $\nu(a)$. Moreover, if $|P(\nu(a))| = 2$, then $\nu(a)$ should prefer $\mu(\nu(a)) = \pi(\nu(a))$ to a, as a should be inferior by the definition of a party permutation.[29] ■

6.3 Weak Stability against Robust Deviations

In this section we discuss an alternative, weaker version of our solution concept. 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 8. *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*

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

- *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.

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., $\{a\} \notin \mathscr{P}$.[30] 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).

[30] 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.

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 construct 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 8. 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 8. 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 5. 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 6. 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, $\mu = \mathrm{id}$ is the unique matching that is weakly SaRD up to depth 1.

To see $\mu = \mathrm{id}$ is weakly SaRD up to depth 1, it suffices to check that neither $(\{1,2\}, \nu_1)$ nor $(\{1,2,3,4\}, \nu_2)$ is strongly robust up to depth 1, where $\nu_1 = \{\{1,2\},\{3\},\{4\},\{5\}\}$ and where $\nu_2 = \{\{1,2\},\{3,4\},\{5\}\}$, because all the other deviations are symmetric to either of these two. Indeed, these deviations are not strongly robust: $\nu'_1 = \{\{1\},\{2,3\},\{4\},\{5\}\}$ is a deviation from ν_1 with $\nu'_1(1) \not\succ_1 \mu(1)$, and $\nu'_2 = \{\{1,2\},\{3\},\{4,5\}\}$ is a deviation from ν_2 with $\nu'_2(3) \not\succ_3 \mu(3)$.

To see that no other matching is weakly SaRD up to depth 1, again, it suffices to check ν_1 and ν_2 because all the other ones are equivalent to either of these two. Note that $(\{4,5\}, \nu_3)$ is a deviation both from ν_1 and from ν_2, where $\nu_3 = \{\{1,2\},\{3\},\{4,5\}\}$. The only deviation from ν_3 is $(\nu'_3, \{2,3\})$ with $\nu'_3 = \{\{1\},\{2,3\},\{4,5\}\}$, and both 4 and 5 are strictly better off at ν'_3 than either at ν_1 or at ν_2. That is, $(\{4,5\}, \nu_3)$ is a strongly robust deviation up to depth 1 either from ν_1 or ν_2, and thus, neither ν_1 nor ν_2 is weakly SaRD up to depth 1. □

6.4 Relation to Other Solution Concepts

6.4.1 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 is 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 answer this question, we first observe through the following example that a SaRD matching is not necessarily included in the bargaining set.[31]

Example 7. 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 (and hence weakly SaRD, too) up to depth 1: $(D, \nu) = (\{2, 3\}, \{\{1\}, \{2, 3\}\})$ is the only deviation from μ, and this is not (weakly) 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 $(\{2, 3\}, \nu)$.[32] □

To check the other inclusion, a result by Klijn and Massó (2003) in the marriage problem is helpful. To begin, note that the marriage problem can be embedded into the roommate problem as follows: a roommate problem $(N, \succ)$ is a marriage problem if there exist disjoint $M, W \subseteq N$ such that $M \cup W = N$, $m \succeq_m m'$ for all $m, m' \in M$ and $w \succeq_w w'$ for all $w, w' \in W$. In the marriage problem, Klijn and Massó (2003) call a matching μ *weakly stable* if for any blocking pair $(m, w) \in M \times W$, either [1] there exists $m' \in M$ such that

[31] While there exist a number of different definitions for bargaining sets, the main point of the following example is valid with all of those the authors are aware of, including the ones by Aumann and Maschler (1964), Mas-Colell (1989), and Zhou (1994).

[32] In the standard definitions of bargaining sets, the agents involved in a counterobjection (i.e., $\{1, 3\}$ in this case) are required to get weakly better off than at the original outcome (i.e., μ in this case), not at the objection that they counter (i.e., ν in this case).

$m' \succ_w m$ and (m', w) is a blocking pair for μ, or [2] there exists $w' \in W$ such that $w' \succ_m w$ and (m, w') is a blocking pair for μ. Klijn and Massó (2003, Theorem 4.2) show that in the marriage problem, a matching is in Zhou's (1994) bargaining set if and only if it is weakly stable and weakly Pareto efficient. The next example demonstrates that a weakly stable matching may not be weakly SaRD up to any depth k and consequently, Zhou's (1994) bargaining set is not included in the set of SaRD matchings up to any depth k.

Example 8. Let $N = \{m_1, m_2, w_1, w_2, w_3\}$ and $\succ$ be such that

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

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 both weakly stable and weakly Pareto efficient.[33] By Klijn and Massó (2003, Theorem 4.2), it is thus in Zhou's (1994) bargaining set. However, this μ is neither SaRD nor weakly SaRD up to any depth k. To see this, consider a deviation $(\{m_1, m_2, w_1, w_2\}, \nu)$ from μ where $\nu = \{\{m_1, w_2\}, \{m_2, w_1\}, \{w_3\}\}$. As there is no deviation from ν (i.e., ν is stable), this is a (strongly) robust deviation from μ up to any depth $k \geq 1$. □

Combining the observations in the two examples, we obtain the following:

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

[33] There are four blocking pairs for μ: (m_1, w_1), (m_1, w_2), (m_2, w_1), and (m_2, w_2). Regarding (m_1, w_1), for instance, we have $m_2 \succ_{w_1} m_1$ and (m_2, w_1) being a blocking pair for μ. All the other three pairs are symmetric. Note also that μ is weakly Pareto efficient because no other matching is strictly preferred by w_3.

6.4.2 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 might appear to resemble indirect dominance in the definition of stable sets.[34] In relation to the farsighted stable set, we make two 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.[35] 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 also 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 definitions 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 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.

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

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

6.4.3 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 it satisfies the following property for each $P \in \mathscr{P}$: if $|P|$ is even, $\mu(a) \in \{\sigma(a), \pi(a)\}$ for all $a \in P$; if $|P|$ is odd, $\mu(a) \in \{\sigma(a), \pi(a)\}$ for all $a \in P$ except for a unique $b \in P$ such that $\mu(b) = b$. □

That is, $\mathscr{P}$-stability requires to match as many "adjacent" pairs as possible in both even and odd parties. This is in contrast with our construction of SaRD matchings: in general, more than one members of a same odd party are unmatched at the outcomes of our algorithm in Section 3. However, we can relate the $\mathscr{P}$-stable matchings and our concept of SaRD as follows:

Proposition 5. *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. Suppose $\#(N, \succ) = 2k + 1$ and fix an arbitrary $\mathscr{P}$-stable matching μ'. By definition, for each odd party $P \in \mathscr{P}(\sigma)$, there exists one and only one agent a_P such that $\mu'(a_P) = a_P$. Let Λ_0 to be the set of all such agents, and label its elements as $\Lambda_0 = \{x_1, \ldots, x_T\}$, where T is the number of the odd parties in $\mathscr{P}(\sigma)$. Then we construct μ from μ' and Λ_0 by iterating the following steps:

Step 0: For each $a \notin \Lambda_0$, let $\mu(a) := \mu'(a)$.

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., $y_t \succsim_{x_t} y$ 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 $\mu(a) := a$.

Note that the resulting μ is a regular matching that "includes" the $\mathscr{P}$-stable matching μ'. Moreover, μ also satisfies Properties 1–2 and hence, we can apply all the Lemmas in Section 4.[36]

Now, 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 Claims 1–2. Note that there is $b \in D \cap I_\mu^\circ$ by Lemma 1, and that $P(b)$ is an odd party for $I_\mu^\circ \subseteq \Lambda_0$ by construction. More specifically, $\mu\left(\sigma^{2j-1}(b)\right) = \sigma^{2j}(b)$ holds for each $j = 1, \ldots, \frac{|P(b)|-1}{2}$. Let ℓ be the smallest integer such that $\nu\left(\sigma^{2\ell-1}(b)\right) \neq \sigma^{2\ell}(b)$. Such ℓ must exist because by the assumption of $Ch = \varnothing \neq Cy$, $b \in D \cap I_\mu^\circ$ implies $b \in \nu(Cy)$ and hence, $\pi(b) \in Cy \subseteq D$ must hold. Moreover, $\sigma^{2\ell-1}(b)$ is not a member of D and thus, single at ν; otherwise, again by the assumption of $Ch = \varnothing \neq Cy$, $\sigma^{2\ell}(b) \in \nu(Cy) \subseteq I_\mu^\circ$ should hold, but this contradicts $\mu\left(\sigma^{2\ell-1}(b)\right) = \sigma^{2\ell}(b)$. Given these observation, 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 $\nu_\ell \rhd_{D_\ell} \nu_{\ell-1} \rhd_{D_{\ell-1}} \cdots \rhd_{D_1} \nu$. That is, the deviation (D, ν) is not robust up to depth ℓ. Since $\ell \leq \frac{|P(b)|-1}{2} \leq k$ by definition, we complete the proof. ■

[36] To check Property 1, suppose that a is superior for b and $\mu(b) = b$. If $a \notin \Lambda_0$, then $\mu(a) \in \{\pi(a), \sigma(a)\}$ and hence, $\mu(a) \succ_a b$ holds. Otherwise, the assumptions imply that either $a = x_t$ is matched to y_t, who is the best partner among $\widetilde{\Sigma}_{t-1} \ni b$, or $a = y_t$ is matched to x_t, who is superior for a, at some step t. In either case, a prefers $\mu(a)$ to b.

Acknowledgments

This manuscript supersedes a previous working paper by Kasuya and Tomoeda (2012). We thank Isa Hafalir, Mamoru Kaneko, Michihiro Kandori, 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, and Sydney EconCS Workshop 2019. Hirata gratefully acknowledges financial support from JSPS KAKENHI (#16K17081).

References

ABRAHAM, D. J., P. BIRÓ, AND D. F. MANLOVE (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.

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, 185–203.

BIRÓ, P., E. IÑARRA, AND E. MOLIS (2016): “A new solution concept for the roommate problem: $\mathcal{Q}$-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.

CHUNG, K.-S. (2000): “On the Existence of Stable Roommate Matchings,” *Games and Economic Behavior*, 33, 206–230.

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

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

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, forthcoming.

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. (2009): "Credibility, Efficiency, and Stability: A Theory of Dynamic Matching Markets," *mimeo*.

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.

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.

TAN, J. J. M. AND Y.-C. HSUEH (1995): "A Generalization of the Stable Matching Problem," *Discrete Applied Mathematics*, 59, 87–102.

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.

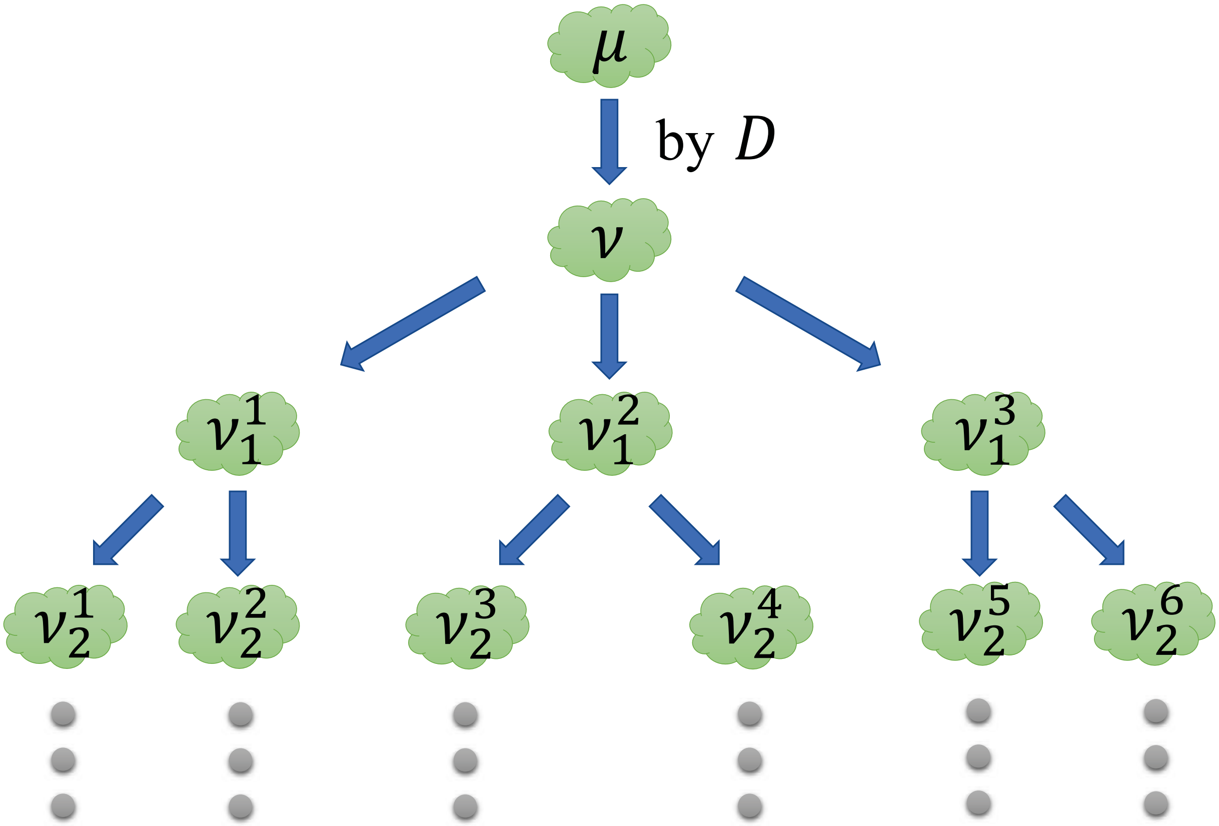
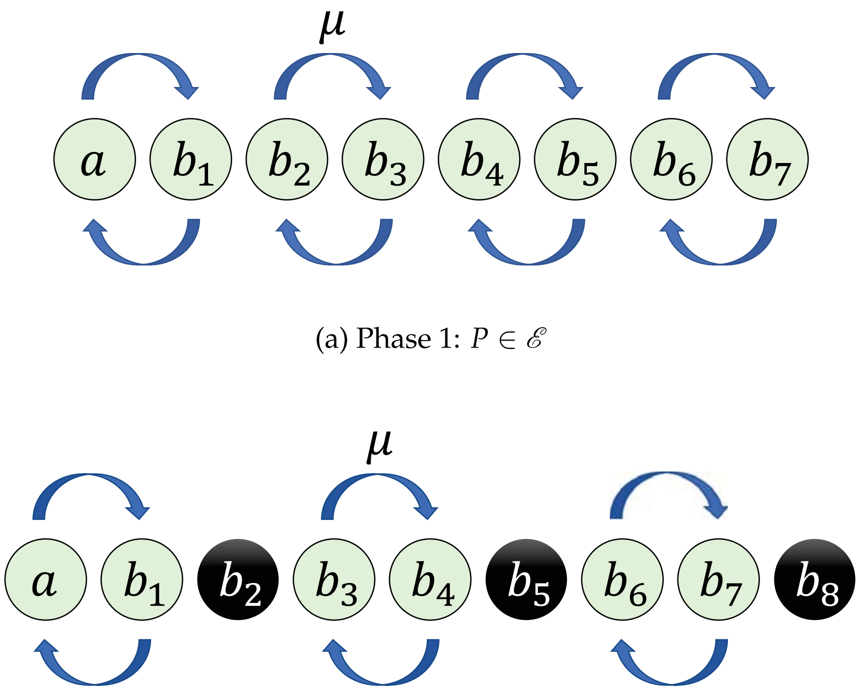


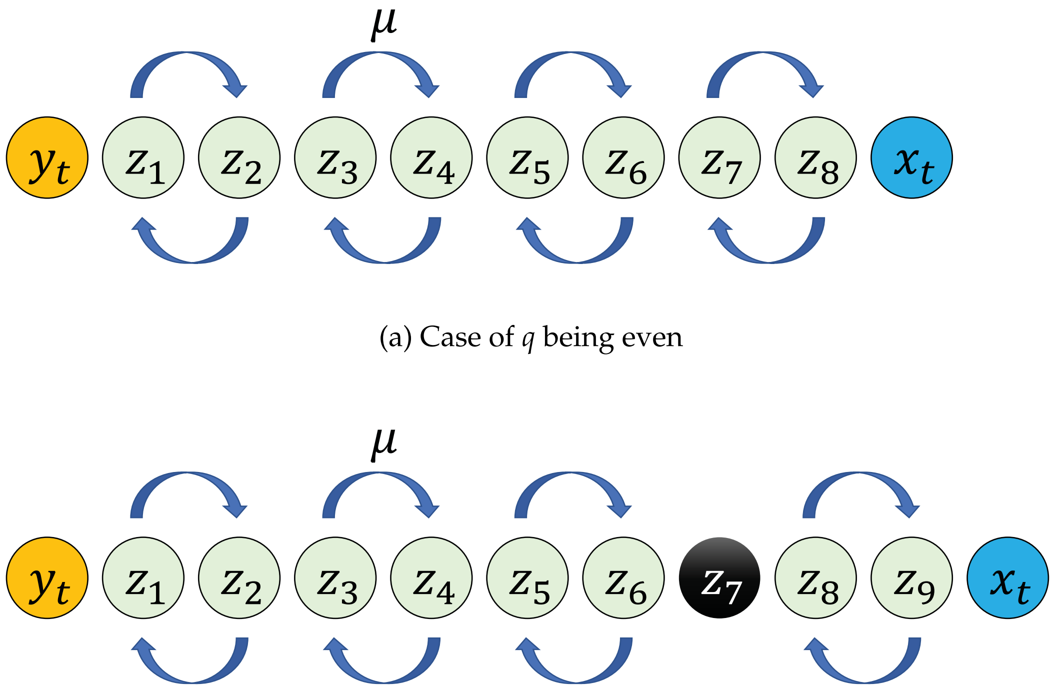
Figure 1: Tree of deviations



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

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

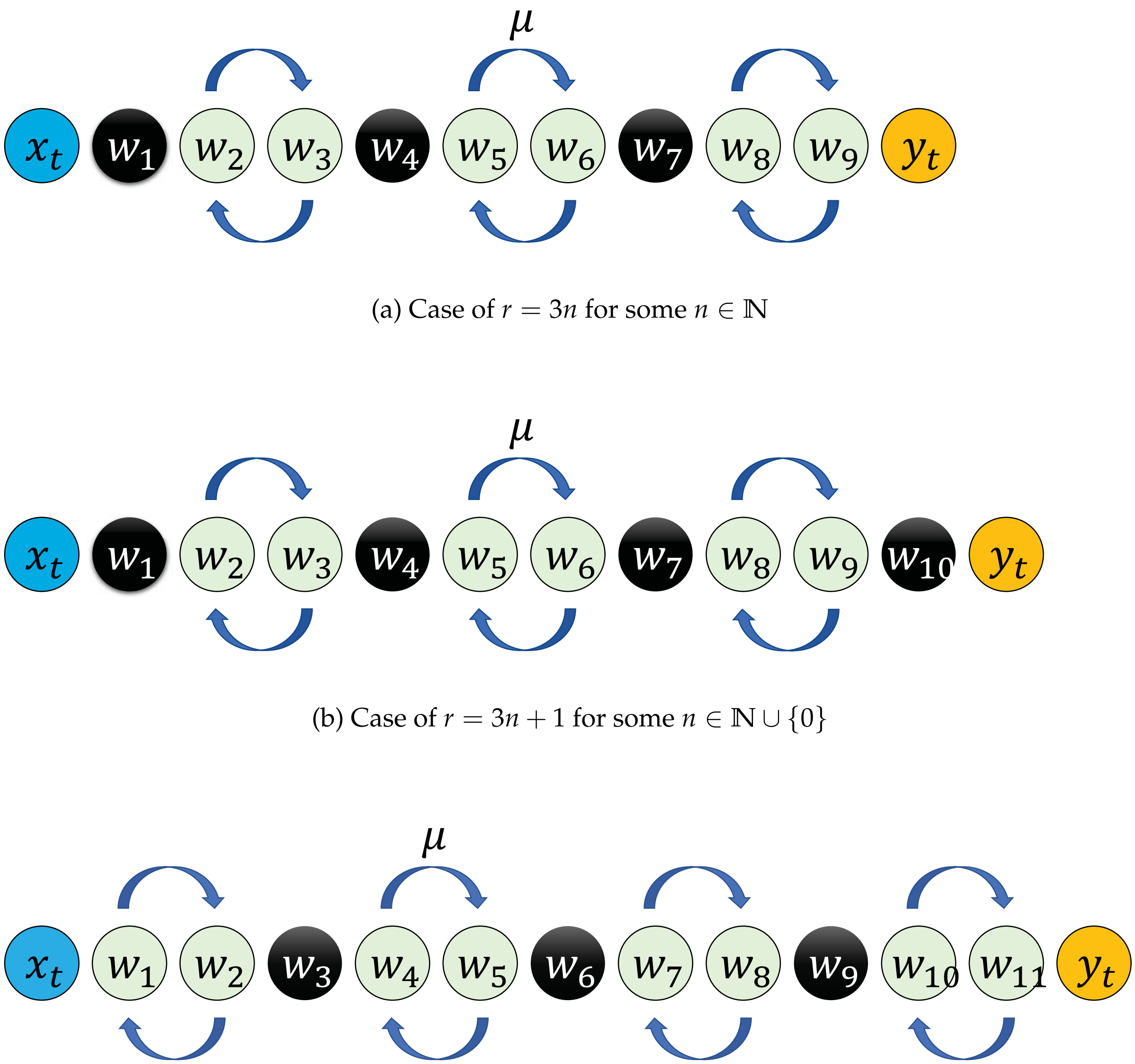
Figure 2: Matching during Phases 1 and 2 of the Algorithm. For each j, b_j represents $\sigma^j(a)$. Each arrow between two agents means they are matched, and the agents represented by black circles are not matched in Phase 2.



(a) Case of q being even

(b) Case of q being odd

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

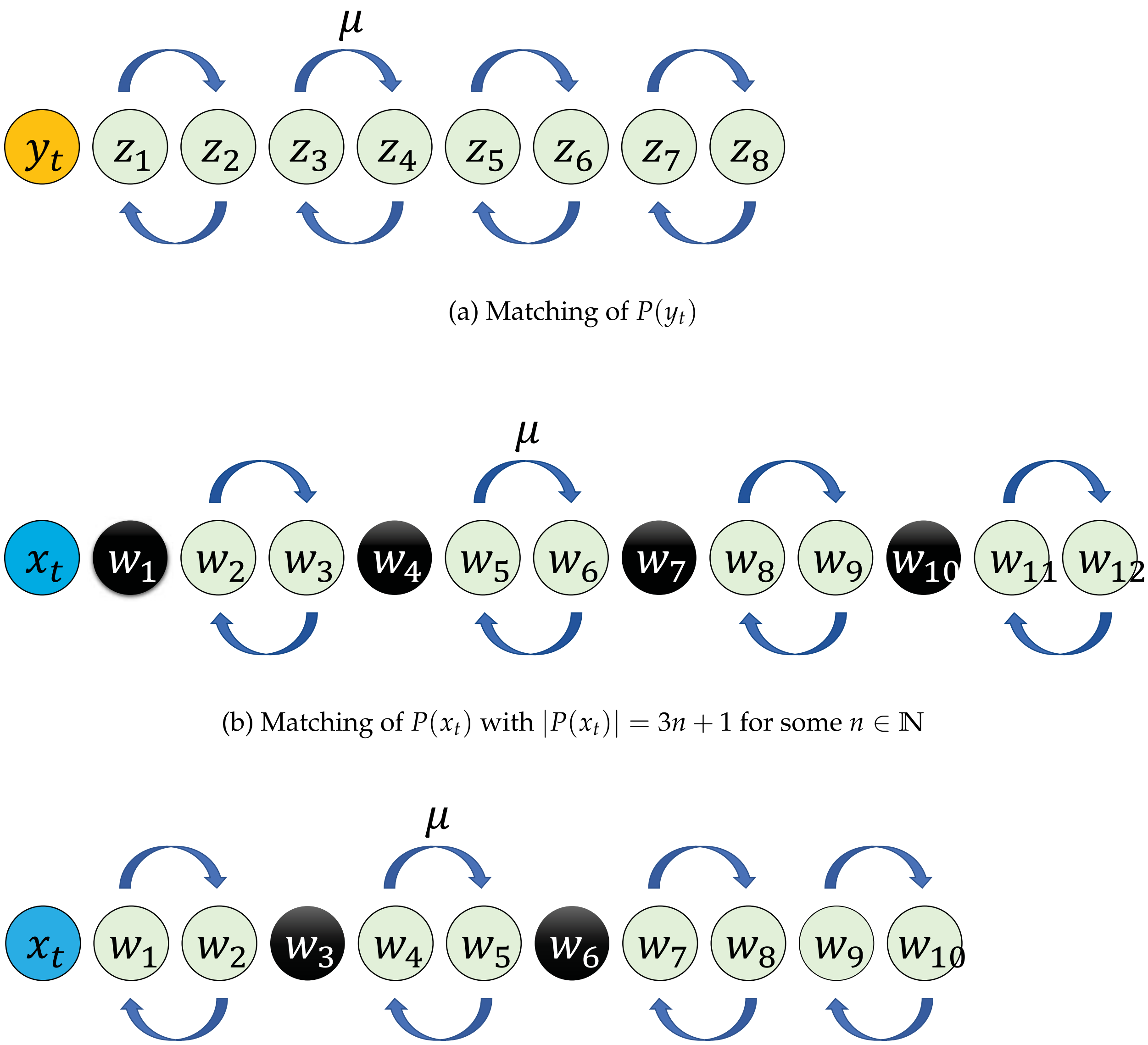


(a) Case of $r = 3n$ for some $n \in \mathbb{N}$

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

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

Figure 4: Matching of the agents among $\sigma(x_t), \ldots, \sigma^r(x_t)$ in Case 1 of Phase 3. For each j, w_j denotes $\sigma^j(x_t)$. Each arrow between two agents means they are matched, and the agents represented by black circles are not matched in this step.

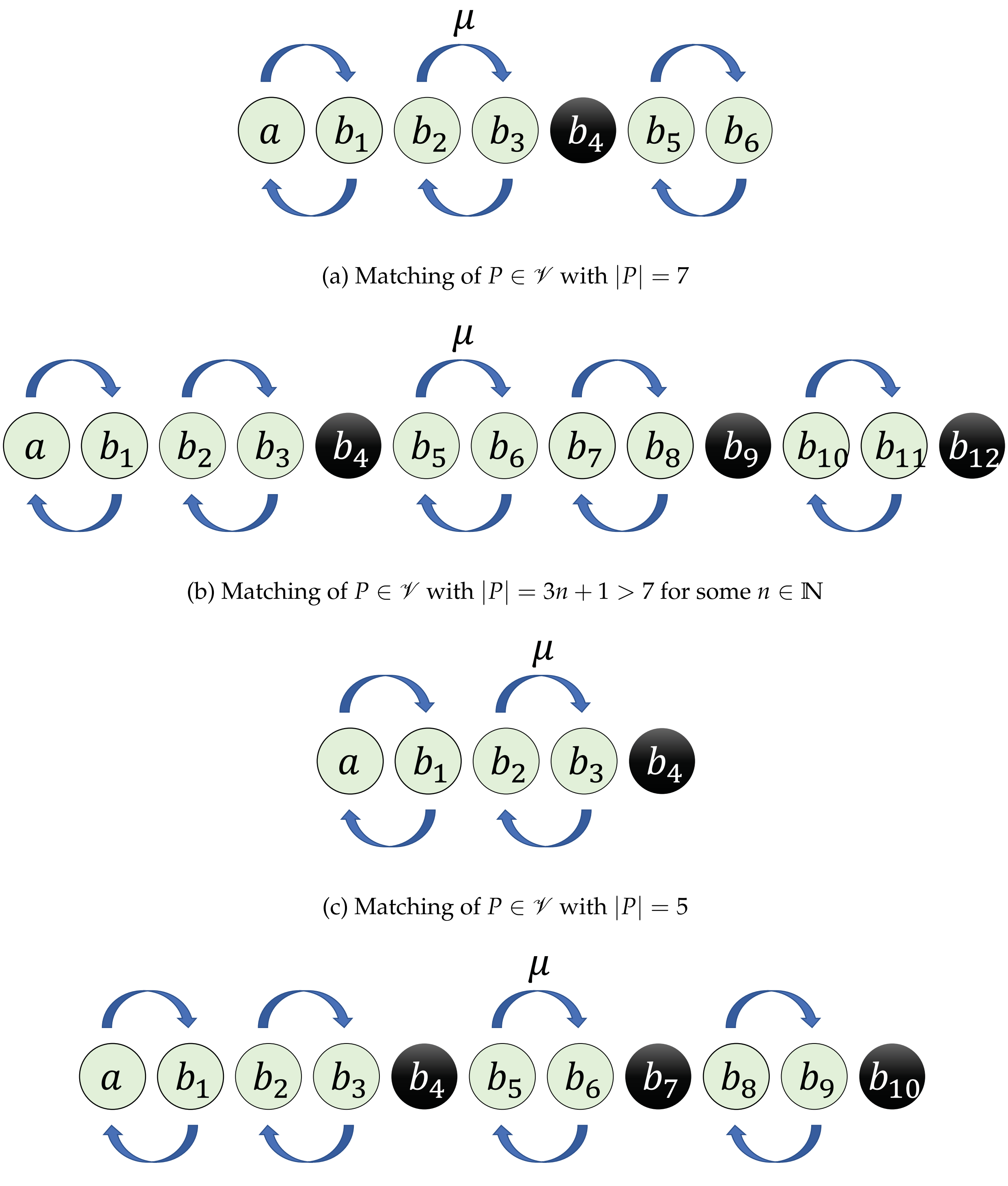


(a) Matching of $P(y_t)$

(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 5: 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.



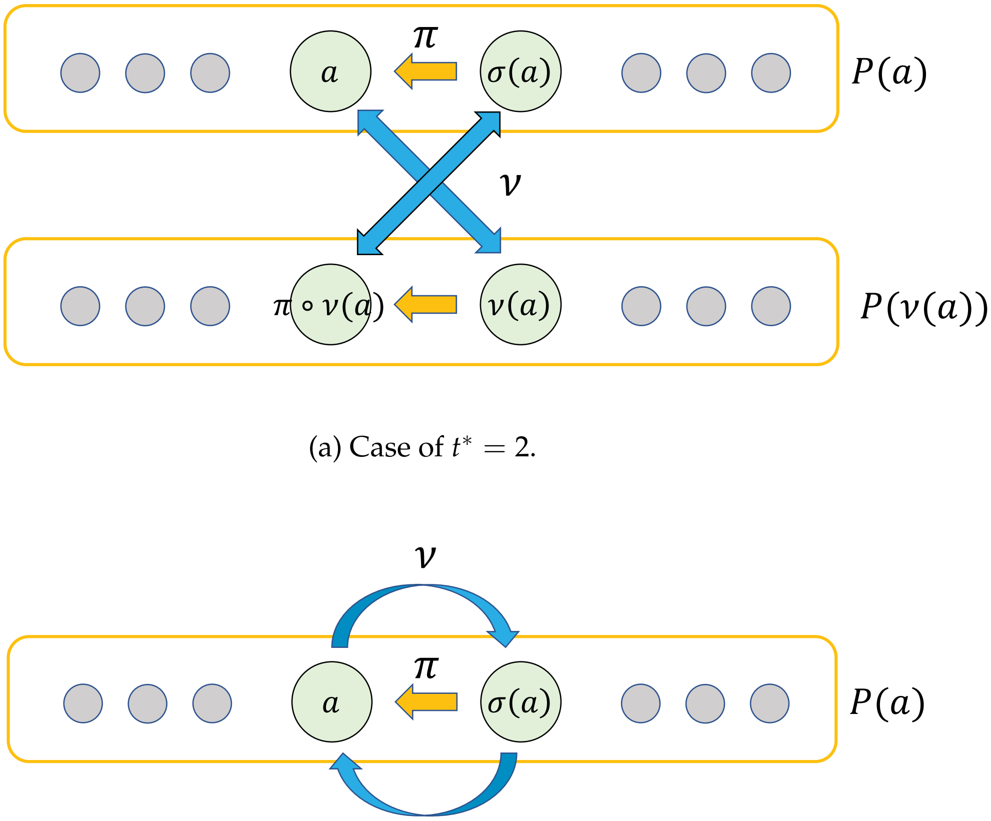
(a) Matching of $P \in \mathscr{V}$ with $|P| = 7$

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

(c) Matching of $P \in \mathscr{V}$ with $|P| = 5$

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

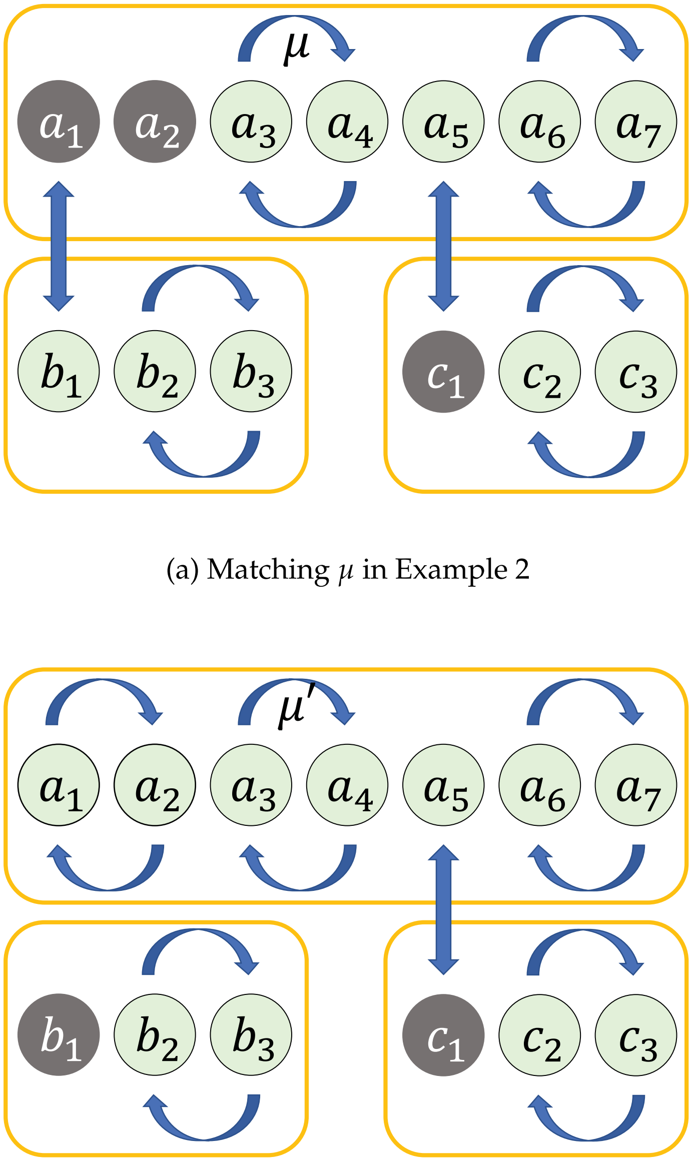
Figure 6: Matching during Phase 4. For each j, b_j denotes $\sigma^j(a)$. Each arrow between two agents means they are matched, and the agents represented by black circles are not matched in this Phase.



(a) Case of $t^* = 2$.

(b) Case of $t^* = 1$.

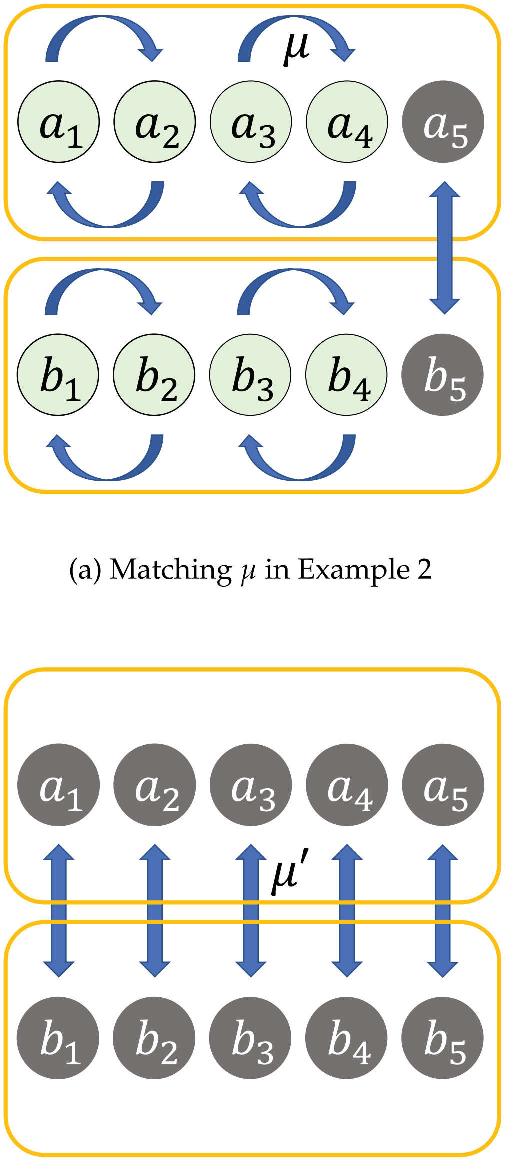
Figure 7: Definition of Cy: If $a \in Cy$, there exists t^* such that $(\pi \circ \nu)^{t^*}(a) = a$.


μ
a_1 a_2 a_3 a_4 a_5 a_6 a_7
b_1 b_2 b_3
c_1 c_2 c_3
μ'
a_1 a_2 a_3 a_4 a_5 a_6 a_7
b_1 b_2 b_3
c_1 c_2 c_3

(a) Matching μ in Example 2

(b) Matching μ' in Example 2

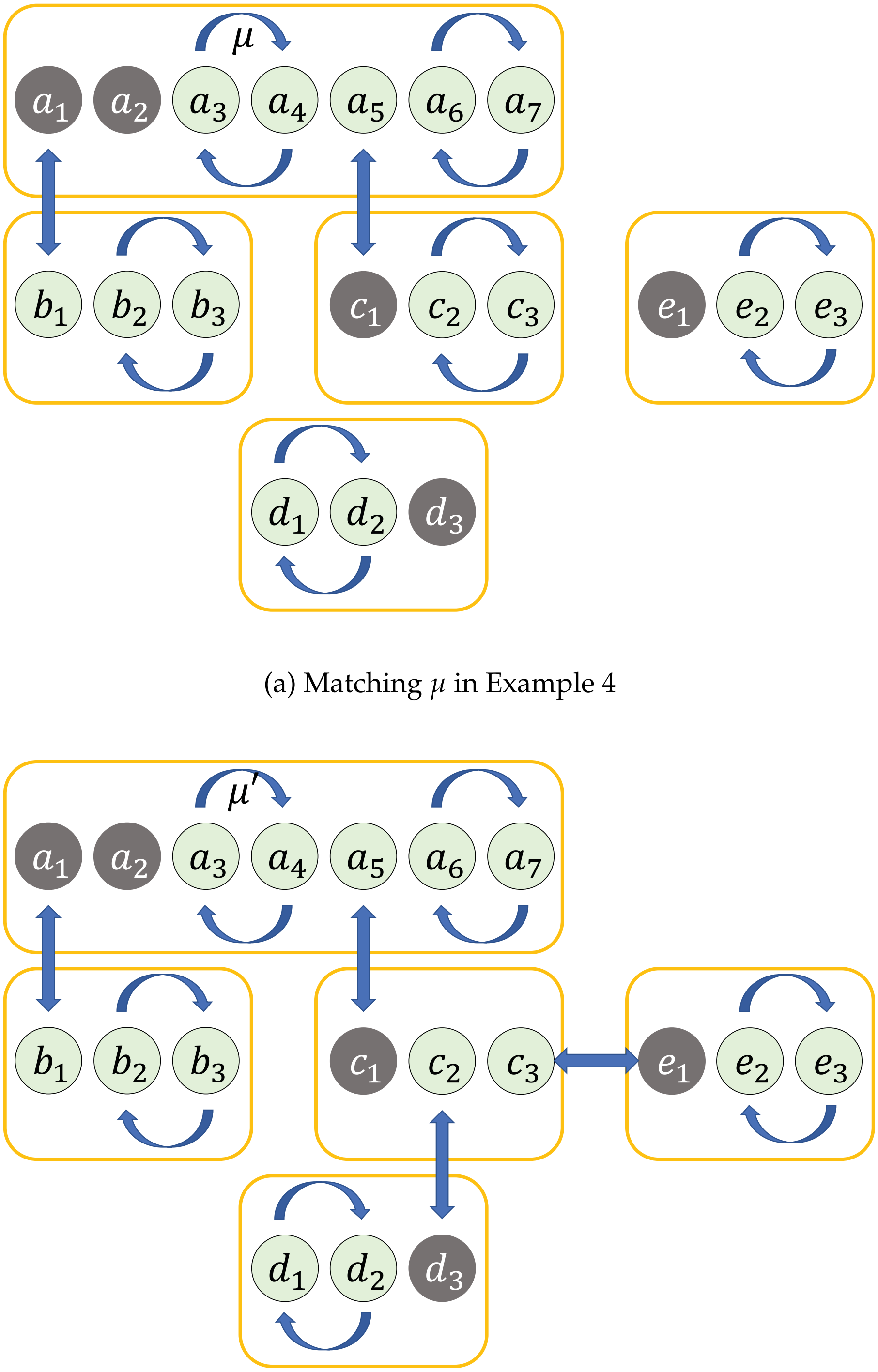
Figure 8: Matching μ and μ' in Example 2. Each box represents an element of the stable partition $\mathscr{P}(\sigma)$. Each arrow between two agents means they are matched, and the agents represented by dark-gray circles are the members of I°_{μ} and $I^{\circ}_{\mu'}$.



(a) Matching μ in Example 2

(b) Matching μ' in Example 2

Figure 9: Matching μ and μ' in Example 3. Each box represents an element of the stable partition $\mathscr{P}(\sigma)$. Each arrow between two agents means they are matched, and the agents represented by dark-gray circles are the members of I°_μ and $I^\circ_{\mu'}$.



(a) Matching μ in Example 4

(b) Matching μ' in Example 4

Figure 10: Matching μ and μ' in Example 4. Each box represents an element of the stable partition $\mathscr{P}(\sigma)$. Each arrow between two agents means they are matched, and the agents represented by dark-gray circles are the members of I°_{μ} and $I^{\circ}_{\mu'}$.