On the Gale-Shapley Algorithm for Stable Matchings with a Partial Honesty Nash Refinement

James P. Bailey
Rensselaer Polytechnic Institute
Troy, NY, United States
bailej6@rpi.edu

Craig A. Tovey
Georgia Institute of Technology
Atlanta, GA, United States
cat@gatech.edu

ABSTRACT

It has long been known that every individually rational matching is obtainable by some Nash equilibrium — even those that make little sense in practice. In the social choice and voting literature, Nash refinements are commonly used to avoid these spurious equilibria. In this paper, we examine the Gale-Shapley algorithm (*deferred acceptance*) where agents behave strategically but are minimally dishonest, a common refinement in the social choice and voting literature. Under this condition we show that when men propose, every equilibrium corresponds to the woman-optimal marriage, thereby yielding a unique prediction for the outcome for the stable matching problem.

KEYWORDS

Matching Market, Stable Matching, Gale-Shapley, Deferred Acceptance, Minimal Dishonesty, Partial Honesty

ACM Reference Format:
James P. Bailey and Craig A. Tovey. 2025. On the Gale-Shapley Algorithm for Stable Matchings with a Partial Honesty Nash Refinement. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 8 pages.

1 INTRODUCTION

The stable matching problem has been applied to a variety of areas including the Nobel Prize winning work by Roth and Shapley on the theory of stable allocations and the practice of market design including the design of the National Resident Matching Program (NRMP) [25, 31] in the United States. Both Canada (CVaRMS) [12] and Japan (JRMP) [22] make use of similar mechanisms. It has also been used for assigning students to schools — both Turkish universities [10] and primary schools in New York [1] and Boston [2]. More recent applications include electrical vehicle recharging [18], ride sharing [11], refugee resettlement [3, 6], and recommendation systems [14].

The stable matching problem seeks to find a matching between two disjoint sets of agents, typically referred to as men and women. In this setting, each agent has strict preferences over the opposite set and we seek a matching that respects both sets of preferences. The standard algorithm for finding a stable matching is the Gale-Shapley algorithm, also known as deferred acceptance [17]. It is also sometimes referred to as the man-optimal (woman-pessimal) algorithm because it always selects the stable matching that is simultaneously best for every man [15] and worst for every woman [23] when men propose.

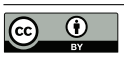
It is well known that every stable matching algorithm is manipulable [30] and agents may misrepresent their preferences to be assigned a preferred partner. The Gale-Shapley algorithm is even known to be manipulable in polynomial time [33]. While there exist stable matching algorithms that are NP-hard to manipulate [28], the Gale-Shapley algorithm is still typically preferred due to both its ease to implement and a key property it has with respect to honesty: when men propose, it is well known that men have no incentive to misrepresent their preferences [13]. This is especially notable in matching markets where one side of the market is incapable of manipulating its preferences. For example, in school admissions, schools' preferences are determined by exam scores — as long as students propose when using the Gale-Shapley algorithm, the resulting mechanism is strategy-proof.

In markets where both sets of agents are strategic, many open problems remain with respect to strategic behavior (see e.g., [20] for a recent review of open problems). In this paper, we revisit a classical analysis of strategic behavior when using the Gale-Shapley algorithm. Under the assumption that all men are honest, Gale and Sotomayor showed that there is always a way for the women to coordinate to obtain the woman-optimal (man-pessimal) stable matching [16]. Further, they showed that the corresponding preferences are a strong Nash equilibrium. Gale and Sotomayor then stated that it would be reasonable to believe that all strong Nash equilibria would be woman-optimal. This would be especially important from a game theory perspective, as ideally equilibria yield unique predictions for the underlying system. However, Gale and Sotomayor expressed disappointment after observing this is not the case:

> *It would have been nice to assert that the [woman-optimal matching] is the only matching obtainable from a strong equilibrium point. Unfortunately, this is not the case.* – Gale and Sotomayor [16].

1.1 Our Contributions

Our paper resolves a significant issue in the use of the Nash equilibrium solution concept in the study of the Gale-Shapley (Deferred Acceptance) algorithm. Specifically, we introduce a common behavioral Nash equilibrium refinement from the social choice and voting literature to study the outcome of the Gale-Shapley algorithm when agents behave strategically and show that every minimally dishonest equilibrium yields the woman-optimal stable matching (Theorem 1) despite the fact that the Gale-Shapley algorithm yields the man-optimal (woman-pessimal) matching when agents are honest.



Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

This allows us to make a precise, unique prediction when agents behave strategically in the setting of stable matchings.

This result notably resolves an issue discovered by both Alcalde and Sönmez [4] and Gale and Sotomayor [16], who showed that the Nash equilibrium and the strong Nash equilibrium solution concepts, respectively, yield non-unique predictions for the Gale-Shapley algorithm. Additionally, our result provides a satisfying response to Gale and Sotomayor's "unfortunate" observation that the strong Nash equilibrium solution concept does not uniquely result in the woman-optimal stable matching.

2 MATHEMATICAL BACKGROUND

An instance of the stable matching problem includes two finite non-empty and disjoint sets of individuals M and W, typically referred to as men and women respectively. Each man $m \in M$ has a strict preference Π_m on $W \cup \{m\}$ where $w_i \Pi_m w_j$ indicates that m prefers to be matched with w_i to w_j. If $m \Pi_m w$ then agent m prefers to be unmatched (equivalently, self-matched) to being matched with w. We remark that some applications require that $w \Pi_m m$ for all $w \in W$ (m always prefers being matched to being unmatched). However, our analysis is consistent with most applications of stable matchings which allow for agents to be unmatched.

For readability, we frequently express the full set of preferences as an ordered list, e.g., $w_1 \Pi_m w_2 \Pi_m m \Pi_m w_3 \Pi_m w_4$ is represented as $(\Pi_m : w_1, w_2, m, w_3, w_4)$. We also denote Π_m^k as the kth element of this list, e.g., in the prior example $\Pi_m^1 = w_1$ and $\Pi_m^3 = m$. Symmetrically, for all $w \in W$, Π_w is a strict preference on $M \cup \{w\}$. A preference profile is given by $\Pi = \bigcup_{i \in M \cup W} \Pi_i$. We denote the set of all possible preferences for agent i as $\mathcal{P}_i$ (for $m \in M$, $\mathcal{P}_m$ corresponds to all permutations of $W \cup \{m\}$) and the set of all possible preference profiles over $M \cup W$ as $\mathcal{P} = \bigtimes_{i \in M \cup W} \mathcal{P}_i$).

A matching μ is a bijection from $M \cup W$ to itself such that $\mu(m) \in W \cup \{m\}$ and $\mu(w) \in M \cup \{w\}$ for each $m \in M$ and $w \in W$. Moreover, the relationship is symmetric; $\mu(i) = j$ if and only if $\mu(j) = i$. We denote $\mu(i)$ as the spouse of i. If $\mu(i) = i$ then i is unmatched (equivalently, self-matched). Individual i strictly prefers μ_1 to μ_2 if and only if $\mu_1(i) \Pi_i \mu_2(i)$.

Stability is a necessary condition for many areas [19, 23, 24]. A stable matching μ represents an equilibrium among agents where no one will leave their partner assigned by μ. Specifically, every agent should be willing to match with their assigned partner and there should be no pair of agents that prefer each other to their assigned spouses. Formally:

Definition 1. *A matching μ is individually rational if for each agent $i \in M \cup W$ either $\mu(i) = i$ (agent i is self-matched) or $\mu(i) \Pi_i i$ (agent i prefers their assigned spouse to being self-matched).*

Definition 2. *A pair $\{m, w\} \in M \times W$ is a blocking pair with respect to μ if $w \Pi_m \mu(m)$ and $m \Pi_w \mu(w)$, i.e., if m and w prefer each other to their respective spouses assigned by μ.*

Definition 3. *The matching μ is* stable *with respect to Π if μ is individually rational and has no blocking pairs with respect to Π. Equivalently, μ is stable if $i \Pi_j \mu(j)$ implies $\mu(i) \Pi_i j$.*

Gale and Shapley [15] proved that a stable matching always exists by developing the Deferred Acceptance (DA) algorithm (more commonly known as the Gale-Shapley algorithm) given below. Informally, each $m \in M$ proposes to their most preferred partner in Π_m that has not yet rejected them. The individual receiving the proposal accepts (thereby becoming matched with m) if either (i) they are unmatched and willing to match with m, or (ii) they are matched to $m' \neq m$ but prefer m to m'. In the second case, the match with m' is broken (m' becomes unmatched). The algorithm concludes once each $m \in M$ is matched or has proposed to every $w \in W$ whom they are willing to match with. Any remaining unmatched $m \in M$ are then self-matched ($\mu(m) = m$).

Algorithm 1 Gale-Shapley Algorithm (Deferred Acceptance)

procedure DA
  $\mu(i) \leftarrow \emptyset$ for $i \in M \cup W$
  $k_m \leftarrow 0$ for $m \in M$
  while there exists $m \in M$ where $\mu(m) = \emptyset$ **do**
    $k_m \leftarrow k_m + 1$
    if $\mu(\Pi_m^{k_m}) = \emptyset$ and $m \Pi_{\Pi_m^{k_m}} \Pi_m^{k_m}$ **then**
      $\mu(m) \leftarrow \Pi_m^{k_m}$ and $\mu(\Pi_m^{k_m}) \leftarrow m$
    else if $m \Pi_{\Pi_m^{k_m}} \mu(\Pi_m^{k_m})$ **then**
      $\mu(\mu(\Pi_m^{k_m})) \leftarrow \emptyset$ and $\mu(\Pi_m^{k_m}) \leftarrow \emptyset$
      $\mu(m) \leftarrow \Pi_m^{k_m}$ and $\mu(\Pi_m^{k_m}) \leftarrow m$
    else
      do nothing
    end if
  end while
  Output μ
end procedure

The Gale-Shapley algorithm has the interesting property that regardless of which order the set of men propose, the same outcome is always obtained [15]. Remarkably, every man obtains his most preferred partner among the set of stable matchings [15]. This matching is referred to as the **man-optimal** stable matching. Formally, the **man-optimal** stable matching μ_M with respect to the preference profile Π is such that for any stable matching μ, for all $m \in M$ either (i) $\mu_M(m) = \mu(m)$ or (ii) $\mu_M(m) \Pi_m \mu(m)$. The **woman-optimal** stable matching is defined symmetrically.

Perhaps more interestingly, the man-optimal stable matching is also the woman-pessimal stable matching [23]; that is, among all stable matchings, the man-optimal stable matching is the least preferred by all women. Symmetrically, the woman-optimal stable matching is the man-pessimal stable matching. This is because the set of stable matching forms a distributive lattice [23] where assigning a woman a better partner results in her original partner obtaining a less preferred spouse.

2.1 The Strategic Stable Matching Game

The ability to select a stable matching relies on the assumption that individuals are truthfully reporting their preferences. However, no stable matching mechanism guarantees strategy-proofness [30]. Individuals might submit a strategic profile $\bar{\Pi}$ that is not equal to the sincere profile Π. To understand the outcome of this strategic behavior, we study a normal-form game with complete information where individuals can submit whichever ordering they like even

though players have common knowledge about the sincere profile Π. Denote this game as the Strategic Stable Matching Game (SSM).

Strategic Stable Matching Game (SSM) with the Gale-Shapley Algorithm

- Each individual has complete information of the sincere preference profile $\Pi = \{\Pi_i\}_{i \in M \cup W}$.
- To play the game, individual i submits strategic preference data $\bar{\Pi}_i \in \mathcal{P}_i$. The collection of all submitted data is denoted $\bar{\Pi}$.
- It is common knowledge that a central decision mechanism will select the man-optimal matching $\mu_{M(\bar{\Pi})}$ with respect to the submitted $\bar{\Pi}$.
- Individual i evaluates $\mu_{M(\bar{\Pi})}$ according to i's partner(s) in the matching(s) $\mu_{M(\bar{\Pi})}$ and i's sincere preferences Π_i.

With respect to the Gale-Shapley algorithm, there are some guarantees with respect to honesty. Dubbins and Freedman showed that when the men propose in the Gale-Shapley algorithm, no man can benefit by altering his submitted preference list [13]. (Roth independently showed this result one year later in [29]). This result is especially pertinent to the college admissions problem where colleges' preferences are determined by examination scores. In this setting, if the Gale-Shapley algorithm is implemented where students "propose" to colleges, each student's best response is to be honest. Colleges are honest by design, resulting in a truthful mechanism for college admissions.

However, in standard implementations of stable matching problems, both sides (in this case, men and women) are allowed to alter their submitted preference list. This provides some recourse to avoid the man-optimal bias introduced by using the Gale-Shapley algorithm. If all other agents are honest, a woman can alter her preference list to obtain her woman-optimal partner [16]. Woman w can accomplish this by submitting the preference list $\mu_W(w)\bar{\Pi}_w w$, where μ_W is the woman-optimal matching, and $w\bar{\Pi}_w m$ for all $m \in M \setminus \{\mu_W(w)\}$, indicating that she is only willing to match with her woman-optimal partner.

In this paper, we aim to understand which stable matchings can be obtained by the Gale-Shapley algorithm when agents behave strategically. Unfortunately, Alcalde and Sönmez have shown that the set of Nash equilibria of SSM corresponds to the set of individually rational matchings [4, 32] regardless of which deterministic stable matching algorithm is used.

Lemma 1 ([4, 32]). *For any deterministic stable matching algorithm, μ is obtained at a Nash equilibrium of SSM if and only if μ is individually rational.*

The result by Alcalde and Sönmez is discouraging. First, it suggests the Nash equilibrium solution concept has little predictive power as there can be an exponential number of stable matchings [23] (although counting the exact number is #P-complete [21]). Second, the Nash equilibria described in Lemma 1 are rather unnatural. The standard proof starts with an arbitrary individually rational matching μ. Then, to construct an equilibrium that yields μ, individual i's submitted preferences are constructed such that i is only willing to match with $\mu(i)$, i.e., $\mu(i)\bar{\Pi}_i i \bar{\Pi}_i v$ for all $v \neq \mu(i)$. We would never expect an individual to lie and indicate they are only willing to match with their least preferred possible partner. Therefore the equilibria described in Lemma 1 are unrealistic and inconsistent with human behavior. In the study of social choice, such unnatural Nash equilibria are removed via Nash equilibrium refinements, i.e., by placing behavioral conditions on agents to only allow "reasonable" behavior.

2.2 Nash Equilibria Refinements

Perhaps the first Nash equilibrium refinement applied to the study of the stable matching game is the strong Nash equilibrium. The concept was first introduced by Aumann [5]. A strong equilibrium point is an equilibrium where no coalition of agents can cooperate in a way in which every agent benefits. The strong equilibrium point is a common solution concept used in social choice, especially in voting theory.

Dubins and Freedman show that no coalition of men can all benefit by misrepresenting their preferences when using the Gale-Shapley algorithm [13] (Roth's result only shows this result for a coalition of size one [29]). Thus, it is generally assumed that all men will be honest when using the Gale-Shapley algorithm. Under this assumption, Gale and Sotomayor then show that there is always a way for women to coordinate to ensure the woman-optimal matching is selected despite the Gale-Shapley algorithm selecting the man-optimal (woman-pessimal) matching when everyone is honest [16]. Further, they show that when men are honest, and women coordinate to obtain the woman-optimal matching, the resulting matching is a strong Nash equilibrium. Given that the women are always able to coordinate to obtain the woman-optimal matching, it would be reasonable to believe that all strong Nash equilibria for SSM with the Gale-Shapley algorithm result in the woman-optimal matching. However, Gale and Sotomayor express disappointment after observing this is not the case.

A criticism of the strong Nash equilibrium is that the requirement is too strong in terms of communication complexity. It requires that agents can engage in an exponential number of private communications, i.e., to verify that a solution is a strong Nash equilibrium, all subsets of $M \cup W$ must communicate and verify they are unable to generate a better solution by altering their preferences. Instead, we focus on another Nash equilibrium refinement in the social choice and voting literature that requires no communication between agents. Instead, it prunes the set of unrealistic equilibria by incorporating studies from behavioral economics that suggest individuals have an aversion to lying.

We consider the minimally dishonest Nash equilibrium [7, 9] concept (equivalently, when there are a finite number of outcomes, truth-bias with distortion costs [26, 27]). Informally, an agent is minimally dishonest as long as being more honest would result in a strictly worse outcome. Equivalently, from a utilitarian perspective, each agent receives a small cost that scales proportional to the size of their lie, therefore, an agent will always prefer submitting a more honest profile as long as the outcome is at least as good for the agent. This Nash equilibrium refinement is logically intuitive, requires no communication between agents, and, most importantly, is supported by various studies in behavioral economics, (see [7] for a thorough discussion).

To establish the "size" of a lie we must first define the distance between two preference profiles. We will use the Kendall Tau (equivalently, bubble sort) distance – the most common way to evaluate the distance between two ordered lists. The Kendall Tau distance counts the number of disagreements between two ordered lists. Formally:

Definition 4 (Kendall Tau Distance). *Let Π_i^1 and Π_i^2 be two preference lists over a set. Then the Kendall Tau distance between Π_i^1 and Π_i^2 is*

$$K(\Pi_i^1, \Pi_i^2) \equiv \left|\left\{\{u, v\} : u\Pi_i^1 v \text{ but } v\Pi_i^2 u\right\}\right|.$$

The formal definition of minimal dishonesty can now be stated succinctly in the notation of Definition 4.

Definition 5 (Minimally Dishonest). *Let Π be the sincere preferences and let $\bar{\Pi}$ be an equilibrium in SSM when using the Gale-Shapley algorithm. Let μ be the stable matching obtained with respect to $\bar{\Pi}$. Agent i is* minimally dishonest *if $K(\bar{\Pi}_i', \Pi_i) < K(\bar{\Pi}_i, \Pi_i)$ implies $\mu(i)\Pi_i\mu'(i)$ where μ' is the matching selected when the profile $\bar{\Pi}' = [\bar{\Pi}_{-i}, \bar{\Pi}']$ is submitted. I.e., submitting a more honest $\bar{\Pi}_i'$ would result in a worse outcome for agent i.*

We remark that when the set of preference profiles is finite, as it is in the stable matching problem, it is straightforward to show that the minimally dishonest Nash equilibrium refinement is equivalent to truth-bias with distortion costs [26, 27] (see [8] for the distinction between the two concepts when the set of outcomes is not finite). In this setting, if agent i is matched with their kth choice Π_i^k by submitting the preference $\bar{\Pi}$, then under the truth-bias concept their associated cost of the matching would be $k + \epsilon \cdot K(\Pi_i, \bar{\Pi}_i)$. Since $K(\Pi_m, \bar{\Pi}_m) \leq \binom{|W\cup\{m\}|}{2}$, $\epsilon < 1/\binom{|W\cup\{m\}|}{2}$ ensures that agent $m \in M$ will always be willing to manipulate their preferences to be matched with a better partner, but will do so in a way that minimizes the size of their manipulation.

Unlike strong equilibria points, a minimally dishonest best response requires no communication between agents. This is a significant improvement in terms of cognitive complexity; when $n = |M| + 1 = |W| + 1$, a minimally dishonest best response requires considering $O(n!)$ more honest preferences whereas strong equilibria points require considering all $O((n!)^n)$ permutations by all agents.

We also consider a cognitively simpler *locally minimally dishonest* best response which only requires agent i to consider swapping two adjacent members of their preference list $\bar{\Pi}_i$. This requires considering at most $n - 1$ profiles that are more honest. Prior to defining a locally minimally dishonest best response, we introduce a notation for swapping two adjacent members of a preference list.

Definition 6. *Suppose that agent i's $k+1$th favorite partner with respect to Π_i is $\Pi_i^{k+1} = u$ and that their kth favorite partner is $\Pi_i^k = v$. Then the new preference obtained by moving u up one position (equivalently, moving v down one position) corresponds to the new preference given by $\bar{\Pi}_i^k = u$, $\bar{\Pi}_i^{k+1} = v$ and $\bar{\Pi}_i^l = \Pi_i^l$ for all $l \notin \{k, k+1\}$.*

The formal definition of local minimal dishonesty can now be stated succinctly in the notation of Definition 6.

Definition 7 (Locally Minimally Dishonest). *Let Π be the sincere preferences and let $\bar{\Pi}$ be an equilibrium in SSM when using the Gale-Shapley algorithm. Let μ be the stable matching obtained with respect to $\bar{\Pi}$. Agent i is locally minimally dishonest if for each u, v where $u\Pi_i v$ but $\bar{\Pi}_i^k = v$ and $\bar{\Pi}_i^{k+1} = u$ for some k, then $\mu(i)\Pi_i\mu'(i)$ where μ' is the matching selected after agent i moves u up one position in $\bar{\Pi}_i$.*

It is trivial to show that every minimally dishonest equilibrium is also a locally minimally dishonest equilibrium. Therefore, when showing properties of minimally dishonest equilibria, it suffices to show the property for locally minimally dishonest equilibria (Section 3). Similarly, to show the existence of locally minimally dishonest equilibria, it suffices to show a minimally dishonest equilibrium exists (Section 4).

We also remark that all of our results also hold when using the Spearman Footrule distance, another common method to measure the distance between two ordinal lists. Like the Kendall Tau distance, it is straightforward to show that the Spearman Footrule distance decreases when switching the order of two adjacent, incorrectly ordered elements of a list. Since local minimal dishonesty only considers such swaps, our results hold for both metrics.

2.3 Movement of Agents in a Preference List

We begin by showing several important properties of moving an agent up one position in another agent's preference list. The first two properties are trivial consequences of Definition 6. The third property describes how the set of stable matchings changes when moving a partner up a single position in a preference list. The result is relatively straightforward and likely has been observed many times in prior works. Nonetheless, the properties will be frequently used in our main results and we formally establish them first to simplify explanations in subsequent sections.

Property 1 (Mostly Preserved Preferences). *Suppose that agent i's $k+1$th favorite partner with respect to Π_i is $\Pi_i^{k+1} = u$ and that their kth favorite partner is $\Pi_i^k = v$. Suppose that $\bar{\Pi}_i$ is obtained when i moves u up one position. Then for all $\{j, k\} \neq \{u, v\}$, $j\Pi_i k$ if and only if $j\bar{\Pi}_i k$. I.e., swapping the positions of adjacent agents u and v does not impact the relative rankings of any other agents.*

Property 1 follows immediately since $\Pi_i^l = \bar{\Pi}_i^l$ for all $l \notin \{k, k+1\}$ and since $l \notin \{k, k+1\}$ implies $l < k+1$ if and only if $l < k$.

Property 2 (Improved Honesty). *Suppose that agent i's $k+1$th favorite partner with respect to $\bar{\Pi}_i$ is $\bar{\Pi}_i^{k+1} = u$ and that their kth favorite partner is $\bar{\Pi}_i^k = v$. Let $\bar{\Pi}_i'$ be the preference list obtained by moving u up one position. If agent i sincerely prefers agent u to agent v ($u\Pi_i v$), then $\bar{\Pi}'$ is more sincere than $\bar{\Pi}$. Specifically, and formally: $K(\bar{\Pi}_i, \Pi_i) = K(\bar{\Pi}_i', \Pi_i) + 1$.*

This result follows immediately from Property 1; since all other preferences are preserved, restoring the preference relation between u and v decreases the number of disagreements with Π_i by exactly one. Notably, this property implies that every minimally dishonest equilibrium is also a locally minimally dishonest equilibrium since the set of more honest profiles considered by local minimal dishonesty is a subset of those considered by minimal dishonesty.

Next, we show that moving agent u up one position in agent i's preference list can only create new stable matchings that match u to i. Symmetrically, if agent v is moved down in i's preference list then the only stable matchings that can be lost will match agent i to agent v.

Property 3 (Set of Stable Matchings)**.** *Consider any profile $\bar{\Pi}$ and let $\bar{\Pi}'$ be obtained when agent i moves agent u up one position in $\bar{\Pi}_i$ causing agent v to move down one position. Let $\mathcal{U}$ and $\mathcal{U}'$ be the set of stable matchings stable with respect to $\bar{\Pi}$ and $\bar{\Pi}'$ respectively. Then:*

(1) if $\mu \in \mathcal{U}' \setminus \mathcal{U}$, then $\mu(i) = u$
(2) if $\mu \in \mathcal{U} \setminus \mathcal{U}'$, then $\mu(i) = v$,

i.e., the movement only creates new stable matchings that match i to u and only removes stable matchings that match i to v.

Proof. The second claim follows immediately from the first claim when considering moving agent v up one position in $\bar{\Pi}'$ causing agent u to move down one position. It now suffices to show the first claim and assume that $\mu \in \mathcal{U}' \setminus \mathcal{U}$. We then consider the cases where $\mu \notin \mathcal{U}$ because μ is not individually rational or μ has a blocking pair.

First we show that if μ is not individually rational with respect to $\bar{\Pi}$, then $\mu(i) = u$ (and $v = i$). Suppose μ is not individually rational with respect to $\bar{\Pi}$ and there is some agent j where $j\bar{\Pi}\mu(j)$ (j prefers to be self-matched). However, with respect to $\bar{\Pi}'$, μ is stable and therefore individually rational implying $\mu(k)\bar{\Pi}'_k k$ for all k. Therefore $\bar{\Pi}'_j$ and $\bar{\Pi}_j$ differ with respect to the pair j and $\mu(j)$. Since only $\bar{\Pi}_i$ changes, $j = i$. Further, by Property 1, the relation between agents with respect to $\bar{\Pi}_i$ and $\bar{\Pi}'_i$ remain unchanged with exception of u and v and $\{u, v\} = \{i, \mu(i)\}$. Since u is moved one position (and ahead of v) in $\bar{\Pi}'_i$, $u\bar{\Pi}'_i v$ and $u = \mu(i)$ and $v = i$ as claimed.

Next, suppose that μ is not stable with respect to $\bar{\Pi}$ because there is a blocking pair $\{j, k\}$ where j and k are distinct. We show that, without loss of generality, that $j = i, \mu(i) = u$, and $k = v$. Since $\{j, k\}$ is a blocking pair, $k\bar{\Pi}_j\mu(j)$ and $j\bar{\Pi}_k\mu(k)$ (they prefer each other to their current match with respect to $\bar{\Pi}$). However, since μ is stable with respect to $\bar{\Pi}'$, either $\mu(j)\bar{\Pi}'_j k$ or $\mu(k)\bar{\Pi}'_k j$ ($\{j, k\}$ is not blocking with respect to $\bar{\Pi}'$). Therefore either j's or k's preferences changed when moving from $\bar{\Pi}$ to $\bar{\Pi}'$. Since only i's preferences change, without loss of generality, $j = i$. Since $k \neq i$, $\bar{\Pi}'_k = \bar{\Pi}_k$. Therefore, in order for $\{i, k\}$ to not be blocking with respect to $\bar{\Pi}'$, $\mu(i)\bar{\Pi}'_i k$ even though $k\bar{\Pi}_i\mu(i)$. By Property 1, the relation between agents with respect to $\bar{\Pi}_i$ and $\bar{\Pi}'_i$ remain unchanged with exception of u and v and $\{u, v\} = \{k, \mu(i)\}$. Since u is moved one position (and ahead of v) in $\bar{\Pi}'_i$, $u\bar{\Pi}'_i v$ and $u = \mu(i)$ and $v = k$ as claimed. □

3 MINIMALLY DISHONEST EQUILIBRIA FOR THE GALE-SHAPLEY ALGORITHM

In this section, we show that when agents are minimally dishonest, there is a unique outcome of the stable matching game — the woman-optimal matching will always be selected when using the man-optimal (Gale-Shapley) algorithm. This is precisely the result that Gale and Sotomayor remarked that they expected when studying strategic behavior when using the Gale-Shapley algorithm [16]. To establish this main result, we first reveal several important properties of a minimally dishonest equilibrium $\bar{\Pi}$.

Lemma 2. *Let $\bar{\Pi}$ be a (locally) minimally dishonest Nash equilibrium for SSM with the Gale-Shapley algorithm that results in the matching μ_M. For $m \in M$, if $\mu_M(m) = \Pi^k_m$ (man m is matched to his kth most preferred partner), then the first k elements of $\bar{\Pi}_m$ are a permutation of the first k elements of Π_m.*

We remark that if we were only considering minimally dishonest best responses, then it would suffice to cite [13, 29] that shows it is always a best response for men to be completely honest when using the Gale-Shapley algorithm, i.e., the only minimally dishonest best response for men is always to be honest. However, the cognitively simpler locally minimally dishonest equilibrium concept requires some additional work to establish the same result in Lemma 6.

Proof of Lemma 2. For contradiction, suppose there exists an agent $v = \Pi^q_m$ where $q > k$ and a $u = \Pi^p_m$ where $p \leq k$ but $v\bar{\Pi}_m u$. Without loss of generality, we may assume that u and v are adjacent in $\bar{\Pi}_m$, i.e., $v = \bar{\Pi}^l_m$ and $u = \bar{\Pi}^{l+1}_m$ for some l: If not, there is either some u' where $u' = \Pi^{p'}_m$ for some $p' \leq k$ where $v\bar{\Pi}_m u'\bar{\Pi}_m u$ or some v' where $v' = \Pi^{q'}_m$ for some $q' > k$ where $v\bar{\Pi}_m v'\bar{\Pi}_m u$. In both cases, we can select u' or v' inductively until obtaining an adjacent pair. We then let $\bar{\Pi}'_m$ be the preferences obtain when m moves u up one position in $\bar{\Pi}_m$ resulting in the new profile $\bar{\Pi}' = [\bar{\Pi}_{-m}, \bar{\Pi}'_m]$.

By Property 2, man m is more honest when submitting $\bar{\Pi}'_m$. Thus, since $\bar{\Pi}_m$ is a minimally dishonest best response, man m must receive a worse outcome when submitting $\bar{\Pi}'_m$, i.e., $\mu_M(m)\Pi_m\mu'_M(m)$ where μ'_M is the matching selected with respect to $\bar{\Pi}'$.

Let $\mathcal{U}$ and $\mathcal{U}'$ be the set of stable matchings stable with respect to $\bar{\Pi}$ and $\bar{\Pi}'$ respectively. First, $\mu_M \in \mathcal{U}'$ since otherwise, by the second part of Property 3, $\mu_M(m) = v = \bar{\Pi}^q_m$ for some $q > k$ contradicting that m is matched with their kth preferred partner. Similarly, by the first part of Property 3, either $\mu'_M(m) = u$ or $\mu'_M \in \mathcal{U}$. If $\mu'_M(m) = u = \Pi^p_m$ for some $p \leq k$, then we contradict that $\mu_M(m)\Pi_m\mu'_M(m)$. Thus both μ and μ' are in both $\mathcal{U}$ and $\mathcal{U}'$.

Since $\mu_M(m)\Pi_m\mu'_M(m)$, $\mu'_M(m) \neq \mu_M(m)$. Since μ_M and μ'_M are the man-optimal matchings with respect to $\bar{\Pi}$ and $\bar{\Pi}'$ respectively, $\mu_M(m)\bar{\Pi}_m\mu'_M(m)$ but $\mu'_M(m)\bar{\Pi}'_m\mu_M(m)$, i.e., the relation between $\mu_M(m)$ and $\mu'_M(m)$ changes for agent m. However, by Property 1, all relations are preserved except the relation between u and v and $(\mu_M(m), \mu'_M(m)) = (u, v)$ since $\mu_M(m)\Pi_m\mu'_M(m)$. Therefore, $u\bar{\Pi}_m v$ and $v\bar{\Pi}'_m u$.

However, this contradicts our selection of u and v. They were selected such that $v\bar{\Pi}_m u$ and $u\bar{\Pi}'_m v$. Therefore man m submitted preferences is a permutation of his k favorite possible partners. □

Next, we show the same result for the set of women. The proof follows similarly to Lemma 2 but changes slightly since the Gale-Shapley algorithm does not select the woman-optimal stable matching with respect to the submitted preferences.

Lemma 3. *Let $\bar{\Pi}$ be a (locally) minimally dishonest Nash equilibrium for SSM with the Gale-Shapley algorithm that results in the matching μ_M. For $w \in W$, if $\mu_M(w) = \Pi^k_w$ (woman w is matched to her kth most preferred partner), then the first k elements of $\bar{\Pi}_w$ are a permutation of the first k elements of Π_w.*

Proof. The first few parts of the proof follow identically to the proof of Lemma 3. For contradiction, we suppose there exists an agent $v = \Pi_w^q$ where $q > k$ and a $u = \Pi_w^p$ where $p \leq k$ but $v\bar{\Pi}_w u$. Without loss of generality, we again may assume that u and v are adjacent in $\bar{\Pi}_w$, i.e., $v = \bar{\Pi}_w^l$ and $u = \bar{\Pi}_w^{l+1}$ for some l and we let $\bar{\Pi}'_w$ be the more honest preference list (by Property 2) obtained after w moves u up one position in her preference list $\bar{\Pi}_w$.

As in the proof of Lemma 2, the matching selected with respect to $\bar{\Pi}' = [\bar{\Pi}_{-w}, \bar{\Pi}'_w]$, μ'_M, must be strictly worse for woman w, i.e., $\mu_M(w)\Pi_w\mu'_M(w)$. Again following identically to Lemma 2, μ_M and μ'_M are both stable with respect to both $\bar{\Pi}$ and $\bar{\Pi}'$.

However, this contradicts that the Gale-Shapley algorithm selects the man-optimal matching. Since men's preferences remain unchanged from $\bar{\Pi}$ to $\bar{\Pi}'$, μ_M is preferred to μ'_M by all men with respect to $\bar{\Pi}$ if and only if μ_M is preferred to μ'_M by all men with respect to $\bar{\Pi}'$. Since μ_M is man-optimal with respect to $\bar{\Pi}$, μ_M is also preferred to μ'_M with respect to $\bar{\Pi}'$ which contradicts that μ'_M is selected by the Gale-Shapley algorithm with respect to $\bar{\Pi}'$. □

Lemma 4. *Let $\bar{\Pi}$ be a (locally) minimally dishonest Nash equilibrium for SSM with the Gale-Shapley algorithm that results in the matching μ_M. Then μ_M is the only matching that is stable with respect to $\bar{\Pi}$.*

Proof. For contradiction, suppose μ_W is the woman-optimal matching with respect to $\bar{\Pi}$ and $\mu_W \neq \mu_M$. Then there is at least one woman w where $\mu_W(w)\bar{\Pi}_w\mu_M(w) = \Pi_w^k$. However, by Lemma 3, $\mu_W(w)\Pi_w\mu_M(w)$ since the first k elements of $\bar{\Pi}_w$ is a permutation of the first k elements of Π_w. This is a contradiction since woman w could alter her preferences to obtain the preferred partner $\mu_W(w)$ by indicating she is only willing to match with $\mu_W(w)$. □

Lemma 5. *Let $\bar{\Pi}$ be a (locally) minimally dishonest Nash equilibrium for SSM with the Gale-Shapley algorithm that results in the matching μ_M. For $i \in M \cup W$, if $\mu_M(i) = \Pi_i^k$ (agent i is matched to her kth most preferred partner), then $\Pi_i^l = \bar{\Pi}_i^l$ for all $l \leq k$. I.e., all agents are honest up to their assigned partner.*

Proof. Similar to the proof of Lemma 2, for contradiction, suppose there exists an agent $v = \Pi_i^q$ where $q \leq k$ and a $u = \Pi_i^p$ where $p < q \leq k$ but $v\bar{\Pi}_i u$. Notably, this means $u\Pi_i\mu_M(i)$. As in Lemma 2, we may assume that u and v are adjacent in $\bar{\Pi}_i$. We then let $\bar{\Pi}'_i$ be the more honest preference list (by Property 2) obtained when i moves u up one position in $\bar{\Pi}_i$ resulting in the new profile $\bar{\Pi}' = [\bar{\Pi}_{-i}, \bar{\Pi}'_i]$.

By Lemma 4, μ_M is the only stable matching with respect to $\bar{\Pi}$. Let μ'_M be the stable matching selected with respect to $\bar{\Pi}'$. By Property 3, if μ is stable with respect to $\bar{\Pi}'$, then either $\mu = \mu_M$ or $\mu(i) = u\Pi_i\mu_M(i)$. Therefore $\mu'_M(i) \in \{\mu_M(i), u\}$, which contradicts local minimal dishonesty since agent i receives at least as good of an outcome when submitting the more honest $\bar{\Pi}'_i$. □

Lemma 6. *Let $\bar{\Pi}$ be a (locally) minimally dishonest Nash equilibrium for SSM with the Gale-Shapley algorithm. For $m \in M$, $\Pi_m = \bar{\Pi}_m$. I.e., all men are honest when selecting the man-optimal matching.*

Proof. Lemma 6 follows from the application of Algorithm 1. Let μ_M be the matching selected with respect to $\bar{\Pi}$ and for man m, let k be such that $\mu_M(m) = \Pi_m^k$. By Lemma 5, $\Pi_m^l = \bar{\Pi}_l$ for all $l \leq k$. Since Algorithm 1 ceases after all men have been matched, and since each man's preference list is accessed by Algorithm 1 in order, the ordering of $\{\bar{\Pi}_m^l\}_{l>k}$ is irrelevant for determining μ_M. Therefore, a (locally) minimally dishonest man m will sincerely report $\{\bar{\Pi}_m^l\}_{l>k}$ since any violation would contradict (local) minimal dishonesty. □

Lemma 7. *Let $\bar{\Pi}$ be a (locally) minimally dishonest Nash equilibrium for SSM with the Gale-Shapley algorithm that results in the matching μ_M. Then μ_M is stable with respect to Π.*

Proof. By Lemma 1, μ_M is individually rational since it is a Nash equilibrium. For contradiction, suppose that μ_M is not stable with respect to Π implying there is a blocking pair $\{m, w\}$. This means that $m\Pi_w\mu_M(w)$ and $w\Pi_m\mu_M(w)$. However, by Lemma 5, each individual is honest up to their partner assigned implying $m\bar{\Pi}_w\mu_M(w)$ and $w\bar{\Pi}_m\mu_M(w)$. Thus, $\{m, w\}$ is also a blocking pair for μ_M with respect to μ_M, contradicting that μ_M is stable with respect to $\bar{\Pi}$. □

Theorem 1. *Let $\bar{\Pi}$ be a (locally) minimally dishonest Nash equilibrium for SSM with the Gale-Shapley algorithm that results in the matching μ_M. Then μ_M is the woman-optimal stable matching with respect to Π.*

Proof. Let μ_W be the sincere woman-optimal stable matching (with respect to Π). We show that μ_W is stable with respect to $\bar{\Pi}$ implying $\mu_M = \mu_W$ since there is a unique stable matching at every equilibrium (Lemma 4). By Lemma 7, μ_M is stable with respect to Π. By [15], for $w \in W$, either $\mu_M(w) = \mu_W(w)$ or $\mu_W(w)\Pi_w\mu_M(w)$ since the woman-optimal matching μ_W assigns each woman her most preferred partner from the set of all stable matchings. By Lemma 5, this implies that each woman is honest up to her woman-optimal partner assigned by μ_W. Similarly, by Lemma 6, each man is completely honest and therefore honest up to his woman-optimal partner. Following the same argument in the proof of Lemma 7, this implies that μ_W is stable with respect to $\bar{\Pi}$ as desired. Thus, $\mu_M = \mu_W$ and the sincere woman-optimal matching is selected when agents are minimally dishonest. □

Theorem 1 recovers the unique prediction that is so important for Nash equilibria — when agents are minimally dishonest while using the Gale-Shapley algorithm, the unique outcome is the woman-optimal stable matching. This is precisely the perverse result that Gale and Sotomayor expected when studying the Gale-Shapley algorithm using the strong Nash equilbrium refinement. Interestingly, we also show that a (locally) minimally dishonest Nash equilibrium is also a strong Nash equilibrium.

Proposition 1. *Let $\bar{\Pi}$ be a (locally) minimally dishonest Nash equilibrium for SSM with the Gale-Shapley algorithm. Then $\bar{\Pi}$ is also a strong Nash equilibrium.*

Proof. Let μ_M be the matching selected with respect to $\bar{\Pi}$. Since, by Lemma 5, every individual is honest up to their partner assigned by μ_M, an individual prefers the matching μ to μ_M with respect to $\bar{\Pi}$ if and only if they also prefer it with respect to Π. Therefore, it suffices to show that no coalition can alter its preferences so that each agent in the coalition prefers the new matching with respect to $\bar{\Pi}$.

First, observe that by Lemma 4, μ_M is the only stable matching with respect to $\bar{\Pi}$ and therefore is the man-optimal matching. By Theorem 1.7.2 of [19], for any coalition that includes at least one man, "it is not possible for the members of the coalition to collectively falsify their preferences so that every one of them obtains a better partner than in" μ_M. Similarly, μ_M is also the woman-optimal stable matching (since it is the only stable matching), and symmetrically by Theorem 1.7.2 of [19] no coalition that includes at least one woman can falsify their preferences so that everyone obtains a strictly better outcome. Thus, every (locally) minimally dishonest Nash equilibrium is also a strong Nash equilibrium. □

4 ON THE EXISTENCE OF MINIMALLY DISHONEST EQUILIBRIA

Our results establish strong properties for minimally dishonest equilibria. However, thus far our results have skimmed over whether minimally dishonest equilibria exist. In this section, we establish that at least one (locally) minimally dishonest equilibrium equilibrium always exists. Moreover, we show that one can be found after applying a polynomial number of best responses.

Lemma 8. *Let Π be a sincere set of preferences, and let $\bar{\Pi}$ be a corresponding submitted preference profile when selecting the man-optimal matching. Suppose $\bar{\Pi}$ is such that:*

(i) There is a unique stable matching μ_M with respect to $\bar{\Pi}$
(ii) Each woman $w \in W$ is honest up to her spouse assigned by μ_M.
(iii) $\Pi_m = \bar{\Pi}_m$ for all $m \in M$ (men are honest)

For each $i \in M \cup W$, let $\bar{\Pi}'_i$ be a minimally dishonest best response to $\bar{\Pi}$ and let $\bar{\Pi}' = [\bar{\Pi}_{-i}, \bar{\Pi}'_i]$. Then

(1) μ_M is the unique stable matching with respect to $\bar{\Pi}'$
(2) If $i \in W$, then i is honest up to her spouse assigned by μ_M.
(3) If $i \in M$, then $\bar{\Pi}'_i = \Pi_i$

Proof. First, we remark that $(i) - (iii)$ imply $\bar{\Pi}$ is a Nash equilibrium (not necessarily minimally dishonest); the result follows identically to the proof of Proposition 1. Claims 1,2 and 3 follow identically to Lemmas 4, 5, and 6, respectively, since the only condition they used was that agents were at a Nash equilibrium where agents are relatively honest, which is guaranteed by conditions $(i) - (iii)$. □

We can then simply apply minimally dishonest best responses iteratively to find a minimally dishonest Nash equilibrium.

Theorem 2. *There exists a Nash equilibrium that can be transformed into a (locally) minimally dishonest Nash equilibrium using at most $|W| \cdot \binom{|M|+1}{2}$ minimally dishonest best responses.*

Proof. We present a Nash equilibrium $\bar{\Pi}$ satisfying properties $(i) - (iii)$ of Lemma 8. If there is an individual $w \in W$ that is not minimally dishonest, the we apply Lemma 8 to obtain a new profile $\bar{\Pi}'$. Notably, since $\bar{\Pi}$ is already a Nash equilibrium, then w only violated minimal dishonesty implying that $\bar{\Pi}'$ is more honest than $\bar{\Pi}$. Further, by Lemma 8, $\bar{\Pi}'$ satisfies properties $(i) - (iii)$ of Lemma 8 and is also a Nash equilibrium. We apply this process iteratively to obtain a minimally dishonest Nash equilibrium. Since each iteration results in a more honest profile, and since honesty is measured with the Kendall Tau distance, a non-negative, integer function, only a finite number of iterations can be applied.

Formally: Let μ_W be the woman-optimal matching with respect to Π. Let $\bar{\Pi}_m = \Pi_m$ for all $m \in M$ and let $\bar{\Pi}_w = \Pi_w$ for all w where $\mu_W(w) = w$ (woman w is unmatched). Finally, for all $w \in W$ where $\mu_W(w) \neq w$, let $\bar{\Pi}_w$ be the profile obtained by truncated Π_w after her partner assigned by μ_W. Formally, if $\mu_W(w) = \Pi^k_w$, then let $\bar{\Pi}^l_w = \Pi^l_w$ for all $l \leq k$, let $\bar{\Pi}^{k+1}_w = w$, and let the remainder of $\bar{\Pi}_w$ be arbitrary.

We denote this preference profile as $\bar{\Pi}^0$ and remark that $\bar{\Pi}^0$ satisfies conditions $(i) - (iii)$ of Lemma 8 and therefore is a Nash equilibrium. Denote the potential function ϕ which measures the total dishonesty of a profile as $\phi(\bar{\Pi}) = \sum_{i \in M \cup W} K(\bar{\Pi}_i, \Pi_i)$. Since all men are honest, $\phi(\bar{\Pi}^0) \leq |W| \cdot \binom{|M|+1}{2}$.

Next, suppose $\bar{\Pi}^l$ satisfies conditions $(i) - (iii)$ of Lemma 8, but is not a minimally dishonest equilibrium. Then there exists a woman $w \in W$ who is not providing a minimally dishonest best response, for by [4], every man's only minimally dishonest best response is to be honest. Let $\bar{\Pi}^{l+1}$ be the new profile obtained when woman w applies her minimally dishonest best response. Since $\bar{\Pi}^l$ is a Nash equilibrium by Lemma 8, $\bar{\Pi}^{l+1}$ is more honest than $\bar{\Pi}^l$ and $\phi(\bar{\Pi}^{l+1}) < \phi(\bar{\Pi}^l)$ Furthermore, by Lemma 8, $\bar{\Pi}^{l+1}$ also satisfies conditions $(i) - (iii)$.

Since ϕ is a non-negative, integer-valued function where $\phi(\bar{\Pi}^0) \leq |W| \cdot \binom{|M|+1}{2}$, this process can be applied at most $|W| \cdot \binom{|M|+1}{2}$ times, i.e., we find a minimally dishonest Nash equilibrium after at most $|W| \cdot \binom{|M|+1}{2}$ iterations. □

5 CONCLUSION

In this paper, we have studied the Gale-Shapley algorithm using a minimal dishonesty refinement to eliminate unrealistic Nash equilibria. We have shown the resulting equilibria always yield the woman-optimal stable matching. From a normative perspective, this result is important as ideally equilibria should yield unique predictions for the underlying system. Further, our results support Gale and Sotomayor's initial belief that the woman-optimal matching will always be obtained when using the Gale-Shapley (man-optimal) algorithm.

ACKNOWLEDGMENTS

Our research has been supported by NSF under grant number CMMI-1335301. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the sponsoring organizations, agencies, or governments.

REFERENCES

[1] Atila Abdulkadiroğlu, Parag A. Pathak, and Alvin E. Roth. 2005. The New York City High School Match. *American Economic Review* 95, 2 (May 2005), 364–367.
[2] Atila Abdulkadiroğlu, Parag A. Pathak, Alvin E. Roth, and Tayfun Sönmez. 2005. The Boston Public School Match. *American Economic Review* 95, 2 (May 2005), 368–371.
[3] Narges Ahani, Tommy Andersson, Alessandro Martinello, Alexander Teytelboym, and Andrew C Trapp. 2021. Placement optimization in refugee resettlement. *Operations Research* 69, 5 (2021), 1468–1486.
[4] Josè Alcalde. 1996. Implementation of stable solutions to marriage problems. *J. Econom. Theory* 69, 1 (1996), 240–254.

[5] Robert J Aumann. 1959. Acceptable points in general cooperative n-person games. *Contributions to the Theory of Games* 4, 40 (1959), 287–324.

[6] Haris Aziz, Jiayin Chen, Serge Gaspers, and Zhaohong Sun. 2018. Stability and Pareto optimality in refugee allocation matchings. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 964–972.

[7] James P Bailey. 2017. *The Price of Deception in Social Choice.* Ph.D. Dissertation. PhD thesis, Georgia Institute of Technology.

[8] James P Bailey and Craig A Tovey. [n.d.]. The Price of Deception in Spatial Social Choice. ([n. d.]).

[9] James P Bailey and Craig A Tovey. 2024. Impact of Tie-Breaking on the Manipulability of Elections. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 105–113.

[10] Michel Balinski and Tayfun Sönmez. 1999. A Tale of Two Mechanisms: Student Placement. *J. Econom. Theory* 84, 1 (January 1999), 73–94.

[11] Siddhartha Banerjee and Ramesh Johari. 2019. Ride sharing. *Sharing Economy: Making Supply Meet Demand* (2019), 73–97.

[12] CVaRMS. [n.d.]. *Canadian Resident Matching Service.* http://www.carms.ca/. http://www.carms.ca/

[13] Lester E. Dubins and David A. Freedman. 1981. Machiavelli and the Gale-Shapley algorithm. *Amer. Math. Monthly* (1981), 485–494.

[14] Farzad Eskandanian and Bamshad Mobasher. 2020. Using stable matching to optimize the balance between accuracy and diversity in recommendation. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 71–79.

[15] David Gale and Lloyd Shapley. 1962. College admissions and the stability of marriage. *Amer. Math. Monthly* (1962), 9–15.

[16] David Gale and Marilda Sotomayor. 1985. Ms. Machiavelli and the stable matching problem. *Amer. Math. Monthly* 92 (1985), 261–268.

[17] David Gale and Marilda Sotomayor. 1985. Some remarks on the stable matching problem. *Discrete Appl. Math.* 11, 3 (1985), 223 – 232.

[18] Enrico Gerding, Sebastian Stein, Valentin Robu, Dengji Zhao, and Nicholas R Jennings. 2013. Two-sided online markets for electric vehicle charging. In *Proc. 12th Int. Conf on Autonomous Agents and Multi-Agent Systems (AAMAS'13)*. Association for Computing Machinery, 989–996.

[19] Dan Gusfield and Robert Irving. 1989. *The Stable Marriage Problem: Structure and Algorithms.* The MIT Press.

[20] Hadi Hosseini and Shraddha Pathak. 2024. Strategic Aspects of Stable Matching Markets: A Survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, Kate Larson (Ed.). International Joint Conferences on Artificial Intelligence Organization, 8077–8085. https://doi.org/10.24963/ijcai.2024/893 Survey Track.

[21] Robert Irving and Paul Leather. 1986. The Complexity of Counting Stable Marriages. *SIAM J. Comput.* 15, 3 (Aug. 1986), 655–667.

[22] JRMP. [n.d.]. *Japan Residency Matching Program.* http://www.jrmp.jp/. http://www.jrmp.jp/

[23] Donald Knuth. 1976. Mariages Stables. *Les presses de L'Université do Montréal* (1976).

[24] David Manlove. 2013. *Algorithmics of Matching Under Preferences.* World Scientific Publishing Co.

[25] NRMP. [n.d.]. *National Residency Matching Program.* http://www.nrmp.org. http://www.nrmp.org

[26] Svetlana Obraztsova and Edith Elkind. 2012. Optimal manipulation of voting rules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26. 2141–2147.

[27] Svetlana Obraztsova, Omer Lev, Evangelos Markakis, Zinovi Rabinovich, and Jeffrey S. Rosenschein. 2017. Distant Truth: Bias Under Vote Distortion Costs. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems* (São Paulo, Brazil). Richland, SC, 885–892. http://dl.acm.org/citation.cfm?id=3091125.3091250

[28] Maria Silvia Pini, Francesca Rossi, K. Brent Venable, and Toby Walsh. 2011. Manipulation complexity and gender neutrality in stable marriage procedures. *Autonomous Agents and Multi-Agent Systems* 22, 1 (2011), 183–199.

[29] Alvin E. Roth. 1982. The economics of matching: stability and incentives. *Math. Oper. Res.* 7 (1982), 617–628.

[30] Alvin E. Roth. 1984. Stability and Polarization of Interests in Job Matching. *Econometrica* 52, 1 (1984), 47–57.

[31] Alvin E Roth and Elliott Peranson. 1999. The redesign of the matching market for American physicians: Some engineering aspects of economic design. *American economic review* 89, 4 (1999), 748–780.

[32] Tayfun Sönmez. 1997. Games of manipulation in marriage problems. *Games Econom. Behav.* 20 (1997), 169–176.

[33] Chung-Piaw Teo and Jay Sethuraman. 1998. The Geometry of Fractional Stable Matchings and its Applications. *Math. Oper. Res.* 23 (1998).