The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Coalition Manipulation of Gale-Shapley Algorithm*

Weiran Shen, Pingzhong Tang
Institute for Interdisciplinary Information Sciences
Tsinghua University
Beijing, China
{emersonswr,kenshinping}@gmail.com

Yuan Deng
Department of Computer Science
Duke University
Durham, NC 27708, USA
ericdy@cs.duke.edu

Abstract

It is well-known that the Gale-Shapley algorithm is not truthful for all agents. Previous studies in this category concentrate on manipulations using incomplete preference lists by a single woman and by the set of all women. Little is known about manipulations by a subset of women.

In this paper, we consider manipulations by any subset of women with arbitrary preferences. We show that a strong Nash equilibrium of the induced manipulation game always exists among the manipulators and the equilibrium outcome is unique and Pareto-dominant. In addition, the set of matchings achievable by manipulations has a lattice structure. We also examine the super-strong Nash equilibrium in the end.

Introduction

The stable matching theory was introduced by Gale and Shapley (1962). Since then, stability has been a central concept in matching market design. The area has attracted intensive research attention, putting theory into practice through a large amount of important applications, such as college admissions and school matchings (Abdulkadiroglu and Sönmez 2003; Abdulkadiroğlu, Pathak, and Roth 2005; Gale and Shapley 1962), hospitals-residents matchings (Irving and Manlove 2009; Irving, Manlove, and Scott 2000; Roth 1996), kidney exchange programs (Abraham, Blum, and Sandholm 2007; Roth, Sönmez, and Ünver 2004; 2005; Liu, Tang, and Fang 2014), and water right trading (Liu et al. 2016; Zhan et al. 2017).

We study the standard stable matching model, where two set of agents, namely men and women, have preferences over each other. A matching is a one-to-one correspondence between the two sets. A pair of a man and a woman, who are not matched together, but prefer each other to their designated partner, is said to be a *blocking pair*. A matching is called *stable* if there exists no blocking pair. The Gale-Shapley algorithm, which was first proposed by Gale and Shapley (1962), takes as input the preference lists of all agents, and computes a stable matching in $O(n^2)$ time. The algorithm simulates the procedure of men proposing to women. At each round of the procedure, each man proposes to his favorite woman among those who have not rejected him yet. Then each woman rejects all men but her favorite one. The algorithm terminates when no man can make any proposal.

It is shown that the matching computed by the Gale-Shapley algorithm is stable and the algorithm is guaranteed to terminate for each legal input, which immediately implies that every instance of the stable matching problem has a stable matching. There are many interesting structural results in the literature of stable matching theory. For example, among all stable matchings, the matching computed by the Gale-Shapley algorithm is preferred by all men to other matchings, and thus is called the M-optimal (W-pessimal) matching. Similarly, the W-optimal (M-pessimal) matching can be found by switching the roles of men and women. In fact, all men and women have opposite preferences over the set of stable matchings, i.e., for every two stable matchings μ_1 and μ_2, all men prefer μ_1 to μ_2, if and only if all women prefer μ_2 to μ_1. Moreover, the set of all stable matchings forms a lattice structure.

Incentive Issue

However, the Gale-Shapley algorithm suffers from the incentive issue, i.e., some agents have incentive to misreport their preference lists. Although it is shown that the Gale-Shapley algorithm is group strategy-proof for all men (Dubins and Freedman 1981)[1], when the algorithm is adopted, the women may have incentives to misreport their preferences. Moreover, a well-known impossibility result by Roth (1982) states that no stable matching algorithm is truthful for all agents.

Gale and Sotomayor (1985) shows that if all women truncate their preference lists properly, the Gale-Shapley algorithm will output a matching that matches each of them to their partner in the W-optimal matching. Teo, Sethuraman, and Tan (2001) provide a polynomial time algorithm to find the optimal single-agent truncation manipulation. However,

*This work was supported in part by the National Natural Science Foundation of China Grant 61561146398 and a China Youth 1000-talent program.

[1]Precisely, *group strategy-proof* means no coalition manipulation can make all men in the coalition strictly better off, in this context. If considering the case where no man is worse off and at least one man is strictly better off, the Gale-Shapley algorithm is not group strategy-proof (Huang 2006).

little is known when only a subset of players can misreport their preference lists.

This paper is directly motivated by the recent reform of the college admissions process in China. In China, all students are required to take the National College Entrance Exam before applying to the universities. The applications are settled by the Ministry of Education using the Gale-Shapley algorithm. However, besides the entrance exam, the Ministry also has the independent admission program (aka the university initiative admission plan). This program allows the universities to conduct independent exams to determine their own ordering of the students. Starting from 2010, these universities began to form leagues and determine their orderings together. Such leagues are widely believed to be beneficial to their members, since they can cooperatively manipulate the admission results. However, these universities are also faced with the problem of competition, since they target for a similar set of students. Such leagues are urged to dissolve by the Ministry for the belief of unfairness.

Our Results

We analyze the manipulation problem in the stable matching problem, where agents can report a preference list over any *subset* of the other sex. Contrary to most existing works, we allow any subset of women to be the manipulators. We show that a strong Nash equilibrium (i.e., no subset of manipulators can deviate and get strictly better off) always exists for any subset of women. Moreover, in the strong Nash equilibrium, each manipulator removes every man below her W-optimal partner on her list and in the induced matching, all manipulators can be matched to their W-optimal partners.

This result generalizes the results by Teo, Sethuraman, and Tan (2001) and Gale and Sotomayor (1985), which consider manipulations by a single woman and the set of all women, respectively. Moreover, the equilibrium outcome is unique and Pareto-dominant for all manipulators, i.e., all manipulators reach a consensus on a single manipulation profile. Furthermore, the set of all stable matchings attainable from general manipulations forms a join-semilattice.

Finally, we show how to check whether such an unique strong Nash equilibrium is a super-strong Nash equilibrium.

Related Works

Knuth, Motwani, and Pittel (1990) show that the number of different partner that a woman can have in all stable matchings is between $\left(\frac{1}{2}-\epsilon\right)\ln n$ and $(1+\epsilon)\ln n$, where n is the number of men and ϵ is a positive constant. Jaramillo, Kayı, and Klijn (2014) study possible manipulations by the women in a many-to-many setting. They consider the so-called dropping strategies, where women are allowed to strategically remove some men in their true preference lists but cannot shuffle their lists. They give an exhaustiveness of this kind of strategies, i.e., for any given stable matching, there exists some dropping strategy that can replicate or improve the matching. Gonczarowski (2014) study group manipulations by all women when the Gale-Shapley algorithm is applied. They also consider dropping strategies and give a tight upper bound on the number of men that must be removed in order for the W-optimal matching to be the final output.

Dworczak (2016) put forward a new matching algorithm to find stable matchings, where all agents are allowed to make proposals. Their algorithm is a natural generalization of the Gale-Shapley algorithm and they also characterize the set of stable matchings by showing that a matching is stable if and only if it is a possible output of their algorithm. Teo, Sethuraman, and Tan (2001) study a different type of manipulation, where a woman can only permute her true preference list[2]. This is a natural constraint when all agents are only allowed to report a complete preference list. They focus on the case where there is only a single manipulator and give an algorithm to find the optimal manipulation that runs in polynomial time. Gupta et al. (2015) extends the algorithm to the so-called P-stable (stable w.r.t preferences P) Nash equilibrium setting.

With the impossibility result by Roth, it is clear that there always exist some agents who have the incentive to manipulate the matching result, no matter what stable matching algorithm is applied. Nevertheless, Pini et al. (2009) design a stable matching mechanism and prove that it is computationally hard to find a manipulation, even for a single manipulator.

Preliminaries

In the standard stable matching problem, there are two sets of agents: the men (denoted by M) and the women (denoted by W). A preference profile P is the collection of the preference lists of all agents. The preference list $P(m)$ of a man $m \in M$ is a strict total order $\succ_m$ over a subset of W, where $w_1 \succ_m w_2$ denotes that m prefers w_1 to w_2. Similarly, the preference list $P(w)$ of a woman $w \in W$ is a strict total order $\succ_w$ over a subset of M. We will use $\succ_m^P$ and $\succ_w^P$ to explicitly refer to the preference lists of m and w in profile P, if multiple preference profiles are considered. However, for simplicity, we always use $\succ_m$ and $\succ_m$ to denote the true preferences of m and w. We slightly abuse notation and use $P(X)$ to denote the preference profile for a set of agents $X \subset M \cup W$.

A matching between men and women is a function $\mu : M \cup W \to M \cup W$, that maps each agent to his or her partner in the matching. For example, $\mu(m) = w$ means that m is matched to w. Thus $\mu(m) = w$ if and only if $\mu(w) = m$. In any matching, a man should be matched with a woman and vice versa. However, we also write $\mu(m) = m$ or $\mu(w) = w$ if m or w is unmatched in μ. We say $\mu_1 \succeq_W \mu_2$, if for all $w \in W$, $\mu_1(w) \succeq_w \mu_2(w)$.

A matching is *individually rational* if no one is matched to someone who is absent from his or her preference list. A pair of man and woman (m, w) is said to *block* a matching μ, if they are not matched together, yet prefer each other to their partners in μ. Such a pair is also called a *blocking pair*. A matching is *stable* if it is individually rational and has no blocking pairs.

[2]This type of manipulations is widely studied in the social choice domain, e.g. (Gibbard 1973).

Recall that the Gale-Shapley algorithm is not truthful for the women (Dubins and Freedman 1981). Let $L \subseteq W$ be the set of manipulators and $N = W \setminus L$ be the set of non-manipulators. We define a manipulation game between the manipulators.

Definition 1 (Manipulation game)**.** *Given the true preference profile of all agents, and a set $L \subseteq W$ of manipulators, a manipulation game is a tuple (L, A^L), where:*

1. *$L \subseteq W$ is the set of manipulators;*
2. *$A^L = \prod_{i \in L} A_i$ is the set of all possible reported preference profiles.*

Remark 1. *Note that in the above definition, all agents in $M \cup N$ are not players. Thus their preference profile is always their true preference profile. The set of all possible preferences A_i depends on different manipulation types which will be defined later, and we only consider the case where all manipulators use the same type of manipulations.*

The outcome of the manipulation game (also called induced matching in this paper) is the matching resulted from the Gale-Shapley algorithm with respect to the reported preference profiles. A manipulator's preference over all possible outcomes of this game is naturally her true preference in P.

We now define two types of manipulations, which determines the elements in A^L.

Definition 2 (General manipulation)**.** *Let $\mathbb{O}_b$ be the set of strict total orders over all possible subsets of M. The manipulators use general manipulations if $A_i = \mathbb{O}_b, \forall i \in L$.*

Definition 3 (Truncation manipulation)**.** *Let $(m_1, m_2, \ldots, m_k)$ be a woman i's true preference list. In truncation manipulations, $A_i = \{(m_1, m_2, \ldots, m_j) \mid \forall j \leqslant k\}$, $\forall i \in L$.*

It is clear that the truncation manipulation is a special case of the general manipulation. The overall preference profile that is taken as input in the Gale-Shapley algorithm is $P = (P(M), P(N), P(L))$. Denote by $S(P)$ the set of all stable matchings under profile P. Moreover, let $S_A(P(M), P(W))$ denote the set of all achievable stable matchings, i.e., stable matchings (with respect to the true preference profile) that can be manipulated to by the manipulators. We sometimes write S_A for short when $(P(M), P(W))$ is clear from the context.

Definition 4 (Nash equilibrium)**.** *A preference profile $P'(L)$ of a manipulation game is a Nash equilibrium if $\forall w \in L$, reporting $P(w)$ results in the best partner for w while assuming the other women reports $P'(L) \setminus \{P(w)\}$.*

In other words, in a Nash equilibrium, any $w \in L$ cannot be matched with a better partner in any stable matching she can manipulate to. We are also interested in the strong notions of Nash equilibrium.

Definition 5 (Strong Nash equilibrium & Super-strong Nash equilibrium)**.** *A Nash equilibrium is* strong*, if no subset of manipulators can jointly manipulate to a matching that is strictly better off for all of them. A Nash equilibrium is* super-strong*, if no subset of manipulators can jointly manipulate to a matching that is weakly better off for all and strictly better off for at least one of them.*

Equivalence between General Manipulation and Truncation Manipulation

Gale and Sotomayor (1985) prove that a strong Nash equilibrium always exists if all women are manipulators and use truncation manipulations. They construct explicitly such a strong equilibrium by letting each woman use a truncation manipulation that removes all men ranked below her W-optimal partner. Ma (2010) also studies truncation manipulations in the same setting and shows that there is only one Nash equilibrium. In addition, the equilibrium profile admits a unique stable matching, namely, the W-optimal matching. Teo, Sethuraman, and Tan (2001) provide a polynomial time algorithm to find the optimal single-agent manipulation. We extend these results to coalition manipulations and consider any subset $L \subseteq W$ as manipulators.

Lemma 1. *Let $P = (P(M), P(W))$ be the true preference profile for all agents. Every matching in $S_A(P)$ induced by a general manipulation can be induced by a truncation manipulation.*

To prove this lemma, we need another useful result from (Gonczarowski and Friedgut 2013):

Theorem 1. *Given agents' strict preferences over agents of the other sex, and a set of manipulators $L \in W$ are allowed to use general manipulations, if no lying woman is worse off, then (1) No woman is worse off; (2) No man is better off.*

Proof of Lemma 1. Let μ be any matching in S_A and $P'(L)$ be the corresponding reported preference profile by the manipulators in a Nash equilibrium of general manipulation. Therefore, μ is the M-optimal matching of $P' = (P(M), P(N), P'(L))$. We construct a truncated preference profile $P_t(L)$ for the manipulators where for each manipulator w, $\mu(w)$ is the last in her preference lists. (If w is single in μ, her preference list remains the same as her true preference list).

Note that μ is stable under the true preference profile. We show that under $P_t = (P(M), P(N), P_t(L))$, μ is also stable. Clearly, μ is individually rational. Assume on the contrary that a pair (m, w) blocks μ under P_t. It follows that $m \succ_w^{P_t} \mu(w)$ and $w \succ_m^{P_t} \mu(m)$. Then we know that $m \succ_w^{P} \mu(w)$ is also true $\forall w \in W$ from the construction of P_t. Also, $w \succ_m^{P} \mu(m)$ is true since no man's preference list is changed. Thus, (m, w) is a blocking pair under P, which contradicts to the stability of μ.

We claim that μ is the W-pessimal matching under P_t. Suppose not and μ^* is the W-pessimal matching and $\mu \neq \mu^*$. Then we let the manipulators manipulate again from P' to P_t and apply Theorem 1 (notice that when applying the theorem, the term "worse off" is with respect to P'). The resulting matching is μ^*. Clearly, no lying woman is worse off according to P'. Thus no woman is worse off. So for any non-manipulator w, we have that $\mu^*(w) \succeq_w^{P'} \mu(w)$. It follows that $\mu^*(w) \succeq_w^{P_t} \mu(w)$, since w is not a manipulator, which contradicts to the assumption that μ^* is the W-pessimal matching under P_t and μ^* is not equal to μ. □

Remark 2. *Note that this result is different from the exhaustiveness result in (Jaramillo, Kayı, and Klijn 2014), since ex-*

haustiveness only requires that $\forall \mu \in S_A(P)$, there exists a truncation manipulation such that the induced matching is weakly preferred *by the manipulators.*

According to Lemma 1, it is therefore without loss of generality to focus on truncation manipulations. In the remainder of this paper, unless explicitly specified, we say a partner or a matching is W-optimal or W-pessimal for a woman if it is so under the true preference profile.

Strong Nash Equilibria

It is well-known that any unmatched woman in a stable matching remains unmatched in all stable matchings.

Theorem 2. *(Roth 1986) Given $P(M)$ and $P(W)$, the set of unmatched agents is the same among all stable matchings.*

Recall that a Nash equilibrium induces a stable matching under true preferences (Gale and Sotomayor 1985). Any unmatched woman in the W-optimal matching has no incentive to misreport since she will always be unmatched. Thus, we only need to consider the case where no manipulator is unmatched in the W-optimal matching.

Lemma 2. *Let $(P(M), P(W))$ be the true preference profile and $P'(L)$ be the reported profile by the manipulators. If for each manipulator w, her W-optimal partner is not removed from her list in truncation manipulation, then $S(P(M), P(N), P'(L)) \subseteq S(P(M), P(W))$.*

Proof. Suppose not and there exists a matching μ which is in $S(P(M), P(N), P'(L))$ but not in $S(P(M), P(W))$. Then, there exists a blocking pair (m, w) in μ under true preference lists. For m, since his preference list is not modified, he prefers w to $\mu(m)$ in true preference lists and the lists after truncation.

If w is not single in μ, then m is still in her preference list after truncation manipulation since $m \succ_w \mu(w)$, which forms a blocking pair in μ with respect to the truncated preference lists. Otherwise, if w is single in μ, notice that, since the order of each man and each woman's preference list is not changed, the W-optimal matching is in $S(P(M), P(N), P'(L))$. Thus, w is single in the W-optimal matching and according to the assumption, she is not a manipulator. □

Theorem 3. *In truncation manipulations, it is a strong Nash equilibrium that each manipulator removes every man below her W-optimal partner on her list. Furthermore, in the induced matching, all manipulators can be matched to their W-optimal partners.*

Our proof is based on the following theorem.

Theorem 4 (Limits on successful manipulation, (Demange, Gale, and Sotomayor 1987))**.** *Let P be the true preferences (not necessarily strict) of the agents, and let P' differ from P in that some coalition C of men and women mis-state their preferences. Then there is no matching μ, stable for P', which is strictly preferred to every stable matching under the true preferences P by all members of C.*

Proof of Theorem 3. We first prove that each manipulator is matched to her W-optimal partner if they report $P'(L)$. According to Lemma 2, the induced matching μ must be in $S(P(M), P(W))$. Notice that the W-optimal matching is still in $S(P(M), P(N), P'(L))$. Thus, according to Theorem 2, each manipulator is not single after manipulation, and she cannot be matched with a man worse than her W-optimal partner since she already removed him. Also, each woman cannot get a partner better than her W-optimal partner. Thus, all manipulators must be matched with their W-optimal partner.

Next we show that it is a strong Nash equilibrium for all manipulators to do so. Note that for all stable matchings in $S(P(M), P(N), P'(L))$, each manipulators are matched with their W-optimal partner. Applying Theorem 4 with $C \subseteq L$, we can conclude that there is no matching in $S(P(M), P(N), P'(L \setminus C), P'(C))$ is strictly preferred to every stable matching in $S(P(M), P(N), P'(L))$ for all members of C. Therefore, the constructed strategy profile is a strong Nash equilibrium. □

This result generalizes the result by (Gale and Sotomayor 1985) and (Teo, Sethuraman, and Tan 2001), which only considers manipulations by the set of all women. If the set of manipulators contains only one woman, the problem becomes a single-agent manipulation and Theorem 3 can also be applied. Thus, in coalition manipulations, every manipulator is matched with the same man as in her best single-agent manipulation.

As Theorem 3 states, there is no conflict between different manipulators in the general manipulation. In fact, each woman can individually perform their optimal singleton manipulation, which provides her W-optimal partner. When combining together, all of them can still be matched to their W-optimal partners.

Lattice Structure

A lattice is mathematical structure containing a set where any two elements have a unique supremum. Formally,

Definition 6 (Join and Meet)**.** *Let $\preceq$ be a partial order defined over a set L for any subset S of L. Let e be an upper bound of S if $e \succeq s$, $\forall s \in S$. e' is a join of S if e' is an upper bound of S and for any upper bound e of S, $e' \preceq e$. Similarly, let e be an lower bound of S if $e \preceq s$, $\forall s \in S$. e' is a meet of S if e' is an lower bound of S and for any lower bound e of S, $e' \succeq e$.*

Definition 7 (Lattice)**.** *L is a join-semilattice if every two-element subset $\{e_1, e_2\} \subseteq L$ has a join. L is a meet-semilattice if every two-element subset $\{e_1, e_2\} \subseteq L$ has a meet. A lattice is both a join-semilattice and a meet-semilattice.*

Now we define two notations that will be useful for later arguments.

Definition 8. *Given two matchings μ and μ', define $\mu_\vee = \mu \vee \mu'$ to be the matching that matches each man to his more preferred partner and each woman to her less preferred partner in μ and μ'. Similarly, we can define $\mu_\wedge = \mu \wedge \mu'$,*

which matches each man to his less preferred partner and each woman to her more preferred partner.

The following theorem states that $\mu_\vee$ and $\mu_\wedge$ are not only well-defined matchings, but also essential to the lattice structure of the set of all stable matchings.

Theorem 5 (Conway's Lattice theorem; (Knuth 1976))**.** *When all preferences are strict, if μ and μ' are stable matchings under preference profile P, then $\mu_\vee = \mu \vee \mu'$ and $\mu_\wedge = \mu \wedge \mu'$ are both matchings. Furthermore, they are both stable under P.*

Therefore, the set of all stable matchings is a lattice with $\succeq_M$ and $\succeq_W$. Let μ^L be a partial matching obtained by restricting the corresponding full matching μ to the set of manipulators and moreover, we call μ an *extension* of μ^L. Let S_A^L be the set of partial matchings obtained by restricting all matchings in S_A to the set of manipulators L.

Before we discuss the lattice structure of the set of S_A, we first prove a lemma about the relation between the length of the preference lists and the induced matching. We say a preference profile $P_1 = (P(M), P(N), P_1(L))$ is shorter than another $P_2 = (P(M), P(N), P_2(L))$ if for each $w \in L$, $|P_1(w)| \leq |P_2(w)|$. In other words, P_1 is shorter than P_2 if all manipulators remove no less men in P_1 than in P_2.

Lemma 3. *Let P_1 and P_2 be two preference profiles and μ_1 and μ_2 be the two corresponding M-optimal matchings. If for each manipulator, her W-optimal partner is in both P_1 and P_2 and P_1 is shorter than P_2, then $\mu_1 \succeq_W \mu_2$.*

Proof. The lemma is a corollary of Lemma 2. Since P_1 is shorter than P_2, P_1 can be viewed as a manipulation starting from P_2. By Lemma 2, the set of stable matchings under P_1 is a subset of that under P_2. Moreover, the Gale-Shapley outputs the W-pessimal matching and thus, $\mu_1 \succeq_W \mu_2$. □

Lemma 4. *Given two truncated preference profiles P_1 and P_2 from $(P(M), P(W))$, and two corresponding M-optimal matchings by μ_1 and μ_2. Let $P_\cap = P_1 \cap P_2 = (P(M), P(N), P_\cap(L))$ such that*

$$P_\cap(w) = \begin{cases} P_1(w) & \text{if } |P_1(w)| \leq |P_2(w)| \\ P_2(w) & \text{otherwise} \end{cases}$$

for all $w \in L$. Then the M-optimal matching $\mu_\cap$ under $P_\cap$ is exactly $\mu_\wedge = \mu_1 \wedge \mu_2$.

Proof. It is easy to check that $P_\cap(L)$ is a legal profile for the manipulators, i.e., the preference list for each manipulator w can be obtained by truncating her true preference list.

According to Lemma 3, we have $\mu_\cap \succeq_W \mu_1$ and $\mu_\cap \succeq_W \mu_2$, so that from the definition of $\mu_\wedge$, we have $\mu_\cap \succeq_W \mu_\wedge$. To show that $\mu_\wedge$ is identical to $\mu_\cap$, we only need to show that $\mu_\wedge \succeq_W \mu_\cap$. We claim that $\mu_\wedge$ is a stable matching under $P_\cap$, and thus $\mu_\wedge \succeq_W \mu_\cap$ because $\mu_\cap$ is the W-pessimal matching under $P_\cap$.

For each w, $\mu_\wedge(w) \succeq_w \mu_1(w)$, so $\mu_\wedge(w)$ is in $P_1(w)$. Similarly $\mu_\wedge(w)$ is also in $P_2(w)$. Therefore $\mu_\wedge(w)$ is individually rational under $P_\cap$. Assume that $\mu_\wedge$ is not stable under $P_\cap$. Then there must be a blocking pair (m, w) and $m \succ_w \mu_\wedge(w)$ and $w \succ_m \mu_\wedge(m)$ under $P_\cap$. However, the two inequalities also hold in both $P_1(w)$ and $P_2(w)$. Notice that m and w are unmatched in at least one of the two matchings μ_1 and μ_2, otherwise $\mu_\wedge(w) = m$. Thus, (m, w) blocks either μ_1 or μ_2, which produces a contradiction. □

Lemma 4 indicates that S_A is a join-semilattice. However, it is not a meet-semilattice, i.e., there exists a subset in S_A that does not have a meet. Consider the counter-example shown in Table 1 and 2.

m_1	w_1	w_3	w_5	–	–	–
m_2	w_2	w_3	w_6	–	–	–
m_3	w_3	w_4	–	–	–	–
m_4	w_4	w_3	–	–	–	–
m_5	w_5	w_1	–	–	–	–
m_6	w_6	w_2	–	–	–	–

Table 1: Men's preference lists

w_1	m_5	m_1	–	–	–	–
w_2	m_6	m_2	–	–	–	–
w_3	m_4	m_2	m_1	m_3	–	–
w_4	m_3	m_4	–	–	–	–
w_5	m_1	m_5	–	–	–	–
w_6	m_2	m_6	–	–	–	–

Table 2: Women's preference lists

μ_1		μ_2		$\mu_\vee$	
m_1	w_5	m_1	w_1	m_1	w_1
m_2	w_2	m_2	w_6	m_2	w_2
m_3	w_4	m_3	w_4	m_3	w_4
m_4	w_3	m_4	w_3	m_4	w_3
m_5	w_1	m_5	w_5	m_5	w_5
m_6	w_6	m_6	w_2	m_6	w_6

Table 3: Matching results

Suppose $L = \{w_1, w_2\}$. If w_1 alone lies and cuts her list to the one containing only m_5, the induced matching is μ_1 in Table 3. If w_2 alone lies and lists only m_6, we will get μ_2. But the meet of the these two matchings $\mu_\vee$ cannot be induced by only truncating the preference lists of w_1 and w_2.

Nevertheless, we prove that the set of partial matchings S_A^L is a lattice.

Lemma 5. *Given $P_1 = (P(M), P(N), P_1(L))$ and $P_2 = (P(M), P(N), P_2(L))$, suppose μ_1 and μ_2 are the two corresponding M-optimal matchings. Let $\mu_\vee = \mu_1 \vee \mu_2$. Then $\mu_\vee^L$ is in S_A^L.*

Proof. We construct a preference profile $P_\cup$ as follows. For each $w \in L$, she removes all men ranked below $\mu_\vee(w)$ in her true preference list. We prove that the corresponding M-optimal matching $\mu_\cup$ is an extension of $\mu_\vee^L$, i.e., $\mu_\cup^L = \mu_\vee^L$.

Using similar techniques as in the proof of Lemma 4, we conclude that $\mu_\vee$ is a stable matching with respect to preference profile $P_\cup$. For each $w \in L$, since $\mu_\vee(w)$ is the last one

in her preference list and they must be matched to their W-pessimal partner under $P_\cup$, $\mu_\vee(w)$ must be equal to $\mu_\cup(w)$ and $\mu_\vee^L = \mu_\cup^L$. □

Lemma 4 is true when restricted to manipulators. Combining the above two lemmas together, we immediately get:

Theorem 6. *Given the set of manipulators L and the true preference profiles $(P(M), P(W))$, the set of stable matchings that can be induced by general manipulations, $S_A(P(M), P(W))$, is a join-semilattice, and the set of partial matchings $S_A^L(P(M), P(W))$ is a lattice.*

In a finite join-semilattice, every two distinct matchings have a join and every woman is weakly better off in the join than any of the two matchings. Thus, if there exist two distinct matchings resulting from strong Nash equilibria, at least one matching can be improved from the perspective of women. Thus, the matching induced from the strong Nash equilibria is *unique* and *Pareto-dominant*.

Super-strong Nash Equilibrium

However, a super-strong Nash equilibrium does not always exist.

Example 1. *Consider the following preference lists (see Table 4 and 5).*

m_1	w_1	w_3	w_2
m_2	w_2	w_1	$-$
m_3	w_3	$-$	$-$

Table 4: Men's preference lists

w_1	m_2	m_1	$-$
w_2	m_1	m_2	$-$
w_3	m_1	m_3	$-$

Table 5: Women's preference lists

The only stable matching under true preference lists is $\{(m_1, w_1), (m_2, w_2), (m_3, w_3)\}$. Therefore, this matching is the only possible outcome of a super-strong Nash equilibrium. However, consider the following manipulation:

w_1	m_2	$-$	$-$
w_2	m_1	$-$	$-$
w_3	m_3	$-$	$-$

Table 6: Women's new preference lists

After using the manipulation, the only stable matching is $\{(m_1, w_2), (m_2, w_1), (m_3, w_3)\}$, in which w_1 and w_2 are strictly better off while w_3 remains the same.

The intuition behind the construction is that we consider one manipulator w who keeps her W-pessimal partner m and rejects any better proposals. Therefore, it is equivalent to a manipulation game by removing w from W, m from M and it is possible that in the remaining manipulation game, the W-optimal matching is weakly better off than the W-optimal matching in the original game, though it is unstable with respect to true preference lists. Thus, with the help of manipulator w, a coalition can have a further manipulation to make everyone weakly better off and at least one strictly better off.

Algorithm

But notice that a super-strong Nash equilibrium must also be a strong Nash equilibrium. Since there exists a unique strong Nash equilibrium, one can check the existence and compute a super-strong Nash equilibrium by simply checking whether there exists a deviation to an unstable matching from the unique strong Nash equilibrium outcome.

Given true preference lists $P = (P(M), P(N), P(L))$, a strong Nash equilibrium $P'(L)$, and its induced matching μ, if there exists a way to deviate to an unstable matching μ' such that all manipulators are weakly better off and at least one manipulator w^* with $\mu'(w^*) = m^* \succ_{w^*} \mu(w^*)$ is strictly better off. Consider the modified preference lists $P'_{w^*,m^*}(L)$ from $P'(L)$: (1) w^* removes all men ranked below m^* in her true preference lists; (2) $\forall w \in L$ with $w \succ_{m^*} w^*$, removes m^*.

Since no manipulators accept more men in $P'_{w^*,m^*}(L)$ than in $P'(L)$, either all of them are weakly better off or some of them become unmatched.

Theorem 7. *Let $P = (P(M), P(N), P(L))$ be the true preference profile. Let $P'(L)$ be a strong Nash equilibrium, and μ be the induced matching by $(P(M), P(N), P'(L))$. $P'(L)$ is a super-strong Nash equilibrium if and only if for all $w \in L$ and $m \in P(w)$ with $m \succ_w \mu(w)$, there exists one manipulator $w' \in L$ such that w' becomes unmatched under preference lists $(P(M), P(N), P'_{w,m}(L))$.*

Proof. By Lemma 1, it is without loss of generality to assume that all manipulators only use truncation manipulations.

"only if" direction: Suppose not and for the sake of contradiction, suppose that there exists $w^* \in L$ and $m^* \in P(w^*)$ with $m^* \succ_{w^*} \mu(w)$ such that under preference lists $(P(M), P(N), P'_{w^*,m^*}(L))$, no $w' \in L$ becomes unmatched. Note that no manipulators accept more men in $P'_{w^*,m^*}(L)$ than in $P'(L)$. Therefore, all manipulators are weakly better off since all of them are still matched after deviation. Moreover, w^* can only be matched to a man $m' \succeq_{w^*} m^* \succ_{w^*} \mu(w^*)$ and thus w^* is strictly better off. Thus, we obtain a deviation such that all manipulators are weakly better off and at least one manipulator is strictly better off, implying that $P'(L)$ is not a super-strong Nash equilibrium.

"if" direction: Suppose not and for the sake of contradiction, assume there still exists a deviation to $P''(L)$ with induced matching μ' such that all manipulators are weakly better off and at least one manipulator w^* is strictly better off. Note that it is without loss of generality to assume that each manipulator w removes every man other than $\mu'(w)$ in her lists since it does not change the induced matching. Therefore, no manipulators accepts more men in $P''(L)$ than in

$P'_{w^*,\mu'(w^*)}(L)$. However, we know there exists one manipulator $w' \in L$ such that w' becomes unmatched under preference lists $(P(M), P(N), P'_{w^*,m^*}(L))$. By removing more men from manipulators' preference lists can only cause more manipulators become unmatched. Therefore, $P''(L)$ cannot be a deviation such that all manipulators are weakly better off and at least one manipulator w^* is strictly better off. □

By Theorem 7, we can design an algorithm to compute a super-strong Nash equilibrium by first computing the unique strong Nash equilibrium $P'(L)$ according to Theorem 3. After that, check whether $P'(L)$ is a super-strong Nash equilibrium by enumerating all possible pair of (w, m) with $w \in L$, $m \in P(w)$, and $m \succ_w \mu(w)$. If assume $|M| = |W| = n$. then the total time complexity is $O(n^4)$: there are $O(n^2)$ possible pairs of (w, m) to enumerate and we need $O(n^2)$ time for running Gale-Shapley algorithm on the $(P(M), P(N), P'_{w,m}(L))$.

Future Research

We show that Gale-Shapley algorithm is vulnerable to manipulation when women are allowed to report arbitrary preference lists. But is there any practical matching mechanism that is hard to manipulate? Though all matching mechanism can be manipulated due to the impossibility result Roth (1982), we can still hope for a practical mechanism that is computationally hard to manipulate. It is also interesting to investigate other types of manipulations, e.g. permutation manipulations.

References

Abdulkadiroglu, A., and Sönmez, T. 2003. School choice: A mechanism design approach. *The American Economic Review* 93(3):729–747.

Abdulkadiroğlu, A.; Pathak, P. A.; and Roth, A. E. 2005. The new york city high school match. *American Economic Review* 364–367.

Abraham, D. J.; Blum, A.; and Sandholm, T. 2007. Clearing algorithms for barter exchange markets: Enabling nationwide kidney exchanges. In *Proceedings of the 8th ACM conference on Electronic commerce*, 295–304. ACM.

Demange, G.; Gale, D.; and Sotomayor, M. 1987. A further note on the stable matching problem. *Discrete Applied Mathematics* 16(3):217–222.

Dubins, L. E., and Freedman, D. A. 1981. Machiavelli and the Gale-Shapley algorithm. *American mathematical monthly* 485–494.

Dworczak, P. 2016. Deferred acceptance with compensation chains. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, 65–66. ACM.

Gale, D., and Shapley, L. S. 1962. College admissions and the stability of marriage. *The American Mathematical Monthly* 69(1):9–15.

Gale, D., and Sotomayor, M. 1985. Ms. machiavelli and the stable matching problem. *American Mathematical Monthly* 261–268.

Gibbard, A. 1973. Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society* 587–601.

Gonczarowski, Y. A., and Friedgut, E. 2013. Sisterhood in the Gale-Shapley matching algorithm. *The Electronic Journal of Combinatorics* 20(2):P12.

Gonczarowski, Y. A. 2014. Manipulation of stable matchings using minimal blacklists. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, 449–449. New York, NY, USA: ACM.

Gupta, S.; Iwama, K.; and Miyazaki, S. 2015. Stable Nash equilibria in the Gale-Shapley matching game. *arXiv preprint arXiv:1509.04344*.

Huang, C.-C. 2006. Cheating by men in the Gale-Shapley stable matching algorithm. In *Algorithms–ESA 2006*. Springer. 418–431.

Irving, R. W., and Manlove, D. F. 2009. Finding large stable matchings. *Journal of Experimental Algorithmics (JEA)* 14:2.

Irving, R. W.; Manlove, D. F.; and Scott, S. 2000. The hospitals/residents problem with ties. In *Algorithm Theory-SWAT 2000*. Springer. 259–271.

Jaramillo, P.; Kayı, Ç.; and Klijn, F. 2014. On the exhaustiveness of truncation and dropping strategies in many-to-many matching markets. *Social Choice and Welfare* 42(4):793–811.

Knuth, D. E.; Motwani, R.; and Pittel, B. 1990. Stable husbands. In *Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms*, 397–404. Society for Industrial and Applied Mathematics.

Knuth, D. E. 1976. *Mariages Stables*. Les Presses de I'Universite de Montreal.

Liu, Y.; Tang, P.; Xu, T.; and Zheng, H. 2016. Optimizing trading assignments in water right markets. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, 551–557.

Liu, Y.; Tang, P.; and Fang, W. 2014. Internally stable matchings and exchanges. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, 1433–1439.

Ma, J. 2010. The singleton core in the college admissions problem and its application to the national resident matching program (nrmp). *Games and Economic Behavior* 69(1):150–164.

Pini, M. S.; Rossi, F.; Venable, K. B.; and Walsh, T. 2009. Manipulation and gender neutrality in stable marriage procedures. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 665–672. International Foundation for Autonomous Agents and Multiagent Systems.

Roth, A. E.; Sönmez, T.; and Ünver, M. 2004. Kidney exchange. *The Quarterly Journal of Economics* 119(2):457–488.

Roth, A. E.; Sönmez, T.; and Ünver, M. U. 2005. A kidney exchange clearinghouse in new england. *American Economic Review* 376–380.

Roth, A. E. 1982. The economics of matching: Stability and incentives. *Mathematics of operations research* 7(4):617–628.

Roth, A. E. 1986. On the allocation of residents to rural hospitals: a general property of two-sided matching markets. *Econometrica: Journal of the Econometric Society* 425–427.

Roth, A. E. 1996. The national residency matching program as a labor market. *Journal of the American Medical Association* 275(13):1054–1056.

Teo, C.-P.; Sethuraman, J.; and Tan, W.-P. 2001. Gale-Shapley stable marriage problem revisited: Strategic issues and applications. *Management Science* 47(9):1252–1267.

Zhan, W.; Li, Z.; Tang, P.; Liu, Y.; and Xu, T. 2017. Stability of generalized two-sided markets with transaction thresholds. In *Proceedings of the AAMAS*.