

Sisterhood in the Gale-Shapley Matching Algorithm

Yannai A. Gonczarowski

Einstein Institute of Mathematics and
Center for the Study of Rationality
Hebrew University, Jerusalem, Israel

yannai@gonch.name

Ehud Friedgut*

Faculty of Mathematics and Computer Science
Weizmann Institute, Israel

ehud.friedgut@gmail.com

Submitted: Dec 19, 2011; Accepted: Apr 6, 2013; Published: Apr 17, 2013

Mathematics Subject Classifications: 91B68, 91B08, 91B52

Abstract

Lying in order to manipulate the Gale-Shapley matching algorithm has been studied by Dubins and Freedman (1981) and by Gale and Sotomayor (1985), and was shown to be generally more appealing to the proposed-to side (denoted as the women in Gale and Shapley’s seminal paper (1962)) than to the proposing side (denoted as men there). It can also be shown that in the case of lying women, for every woman who is better off due to lying, there exists a man who is worse off.

In this paper, we show that an even stronger dichotomy between the goals of the sexes holds, namely, if no woman is worse off then no man is better off, while a form of sisterhood between the lying and the “innocent” women also holds, namely, if none of the former is worse off, then neither is any of the latter. These results are robust: they generalize to the one-to-many variants of the algorithm and do not require the resulting matching to be stable (i.e. they hold even in out-of-equilibria situations). The machinery we develop in our proofs sheds new light on the structure of lying by women in the Gale-Shapley matching algorithm.

This paper is based upon an undergraduate thesis (2007) by the first author.

Keywords: matching; manipulation; out-of-equilibrium; unstable

1 Background

1.1 The Gale-Shapley Algorithm

In order to standardize the notation used throughout this paper, and for sake of self-containment, let us quickly recap the scenario introduced in [3] and the Gale-Shapley Algorithm introduced there:

*Research supported in part by I.S.F. grant 0398246 and BSF grant 2010247.

Let W and M be equally-sized finite sets of women and men, respectively. Let each member of W (resp. M) have a strict order of preference over the members of M (resp. W).

Definition 1 (Matching). A one-to-one map between W and M is called a *matching*.

Definition 2 (Stability). A matching is said to be *unstable* under the given orders of preference if there exist two matched couples (w, m) and (\tilde{w}, \tilde{m}) such that w prefers \tilde{m} over m and \tilde{m} prefers w over \tilde{w} . A matching that is not unstable is said to be *stable*.

Definition 3 (The Gale-Shapley Algorithm [3]). The following algorithm is henceforth referred to as the *Gale-Shapley algorithm*: The algorithm is divided into steps, to which we refer as *nights*. On each night, each man serenades under the window of the woman he prefers most among all women who have not (yet) rejected him, and then each woman, under whose window more than one man serenades, rejects every man who serenades under her window, except for the man she prefers most among these men. The algorithm stops on a night on which no man is rejected by any woman, and then each woman is matched with whoever has serenaded under her window on this night.

Theorem 4 ([3]). *The Gale-Shapley algorithm stops and yields a stable matching between W and M (in particular, such a matching exists), and no stable matching is better for any man $m \in M$.*

Generalizations of this algorithm, for scenarios including one-to-many matchings, preference lists that do not include all the members of the opposite sex (“blacklisting”), and mismatched quotas, are also presented in [3]. We discuss these variants in Section 3.

1.2 Previous Results

Theorem 4 states that the stable matching given by the Gale-Shapley algorithm is optimal (out of all stable matchings) for each man. A somewhat reverse claim holds regarding the women:

Theorem 5 ([10]). *No stable matching is worse for any woman $w \in W$ than the matching given by the Gale-Shapley algorithm.*

The benefits of the Gale-Shapley matching algorithm for the men are demonstrated even further in [2], where it is shown that no man can get a better match by lying about his preferences (i.e. manipulate the algorithm by declaring a false order of preference, assuming the algorithm is run according to the “true” preferences of all women and all other men) and, moreover, that no nonempty subset of the men can all get better matches by lying in a coordinated fashion.¹

These observations led to the analysis of the profitability of lying by women in [4], where it is proven that if more than one stable matching exists, then at least one woman

¹The interested reader is referred to [1] for an even stronger non-manipulability theorem, and to [7] for the study of lying by men in a probabilistic setting.

can get a better match by lying about her preferences. It is furthermore shown in [12] that in this case, such a woman can get a better match merely by truncating her preference list (i.e. blacklisting a suffix of her preference list; see Section 3.2). However, as observed in [6, p. 65], there appear to be no systematic results on the possibility that women may benefit by permuting their preference lists in some way (without blacklisting any man). The analysis presented in this paper constitutes a contribution in this direction.

It is easy to show that any woman who is better off as a result of someone's lie (whoever that liar or those liars may be and whatever their lie may be) is matched (due to the lie) to someone who is now worse off. (Indeed, if they were both better off, then the original match could not be stable.) In other words, for every woman who is better off, some man is worse off. In the next section, we show that an even stronger dichotomy between the goals of the sexes holds, namely, if no woman is worse off then no man is better off, while a form of sisterhood between the lying and the "innocent" women also holds, namely, if none of the former is worse off, then neither is any of the latter. The machinery we develop in order to prove these results sheds new light on the structure of lying by women in the Gale-Shapley matching algorithm.

2 Monogamous Matchings

2.1 The Theorem

Let W and M be equally-sized finite sets of women and men, respectively. Let each member of these sets be endowed with a strict order of preference regarding the members of the other set. These will be referred to as the *true* preferences, the run of the Gale-Shapley algorithm according to these preferences will be denoted by OA (original algorithm), and the resulting matching will be referred to as the *original* matching.

Assume that a subset of the women, denoted by L (for liars), declare false orders of preference for themselves. Denote by NA the run of the Gale-Shapley algorithm according to these false preferences for the members of L , and according to the real preferences of other members of W (referred to, henceforth, as *innocent*) and of all the members of M . Call the resulting matching the *new* matching.

Definition 6 (Improvement / Worsening). A person $p \in W \cup M$ is said to be *better off* (resp. *worse off*) if p prefers, according to their true order of preference, their match according to the new matching (resp. the original matching) over their match according to the original matching (resp. the new matching).

Let us now phrase our main result for the above conditions.

Theorem 7 (Sisterhood²). *Under the above conditions, if no lying woman is worse off, then:*

²Theorem 7, along with its proof that we present in Section 2.4, was discovered by the first author while participating in a freshman undergraduate course taught by the second author, in 2004.

(a) *No woman is worse off.*

(b) *No man is better off.*

2.2 Proofs for a Special Case

Definition 8 (Personally-Optimal Lie). A woman $l \in L$ is said to be lying in a *personally-optimal* way if, all other orders of preference being the same, there is no other order of preference she can declare that would result in her being even better off, i.e. her being matched with a man she (truly) prefers over the man matched to her by the new matching.

Theorem 7 is easily provable in the special case in which all women in L lie in a personally-optimal way, as it is possible to show that in this case, the new matching is stable under the true preferences, and thus, by Theorems 4 and 5, no woman is worse off, and no man is better off, than under the original matching. Nonetheless, we show in Section 2.3 that in some cases, when liars may coordinate their lies, it may be rational to lie in a non-personally-optimal way, and that in these cases the resulting matching might be unstable under the true preferences.

We note that if we examine a “lying game” between the members of L , where a player’s strategy is a declaration of a specific order of preference for herself, and the utility for each player is determined by the ranking of her new match according to her true order of preference, then in this game, all lies are personally optimal if and only if the set of lies constitutes a Nash equilibrium. Indeed, the general proof we give below holds even when lies need not be personally optimal, i.e. in out-of-equilibria situations.

Roth (private communication, Dec. 2007) suggested the following sketch of a proof to part (a) of Theorem 7:

- If women can do better than to state their true preferences, they can do so by truncating their preferences (i.e. by each of them blacklisting a suffix of her preference list; blacklisting is discussed in Section 3.2).
- Truncating preferences is the opposite of extending preferences (as discussed in [12] in the context of adding new players).
- When any woman extends her preferences, it harms the other women.

While this also proves Theorem 7 in the special case discussed above (in which all women lie in a personally-optimal way), it appears not to prove Theorem 7 itself (i.e. with no additional assumptions), as while it is true that if a woman can do better than to state her true preferences then she can do better by truncating them (and even more so, there always exists a truncation of her true preferences that constitutes a personally-optimal lie for her), it turns out that there may exist some man whom she can secure for herself by submitting false preferences, but not by truncating her true preferences. This is illustrated by an example in the next section.

2.3 When a Lie Need Not be Optimal

As is observed in Section 2.2, Theorem 7 is easily provable if each lying woman lies in a personally-optimal way, or, more generally, if the new matching is stable under everyone's true preferences. Before continuing to the general proof of this theorem, let us first give an example of a scenario with the following properties:

1. No woman can do better by lying alone (while all others tell the truth).
2. In every conspiracy by more than one woman to lie so that none of them is worse off and at least one of them is better off, there exists a woman who does not lie in a personally-optimal way. In other words, it is rational to lie in a non-personally-optimal way.
3. The resulting matching is not stable under the true preferences and can not be achieved by simply truncating women's true preference lists (even if that is allowed).

This example includes four women (w_1, \dots, w_4) and four men (m_1, \dots, m_4). The orders of preference of the women fulfill:

- w_1 : First choice: m_3 , second choice: m_1 .
- w_2 : First choice: m_3 , second choice: m_1 .
- w_3 : Prefers m_2 over m_1 and prefers m_1 over m_3 .
- w_4 : Any order of preference.

The orders of preference of the men fulfill:

Man	1st choice	2nd choice	3rd choice	4th choice
m_1	w_1	w_3	w_2	w_4
m_2	w_2	w_3	any	any
m_3	w_3	w_2	w_1	w_4
m_4	w_1	w_4	any	any

Let us examine OA :

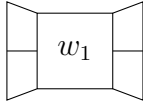
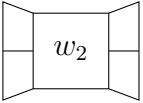
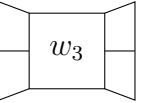
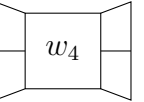
Window	w_1	w_2	w_3	w_4
Night				
1	m_4, m_1	m_2	m_3	
2	m_1	m_2	m_3	m_4

It is clear from this examination that any conspiring subset of the women wishing to alter the outcome (when truncating preference lists is not allowed) has to include w_1 . Also, it may be verified that w_1 cannot lie alone (while all others tell the truth) and become better off. Indeed, in order to become better off, w_1 has to be matched, under the new matching, with m_3 , and since he prefers w_2 over her, this entails his rejection by w_2 , but since he is the first choice of w_2 , this means w_2 has to lie and declare that she prefers another man over him. Let us assume, then, that $L = \{w_1, w_2\}$. (It can be verified that the result would not change even if we admit more women to L .)

It can easily be verified that, given the true orders of preference for everyone except w_1 and w_2 , there exists exactly one combination of false orders of preference for these two women (up to changes in their preferences regarding men who do not reach their windows as long as everyone else tells the truth) that causes them both to be better off:

- w_1 must declare she prefers m_3 over m_4 and m_4 over m_1 .
- w_2 must declare she prefers m_1 over m_3 and m_3 over m_2 .

Let us examine NA under these false orders of preference:

Window \ Night	 w_1	 w_2	 w_3	 w_4
1	m_4, m_1	m_2	m_3	
2	m_4	m_2	m_1, m_3	
3	m_4	m_2, m_3	m_1	
4	m_4	m_3	m_1, m_2	
5	m_4	m_1, m_3	m_2	
6	m_3, m_4	m_1	m_2	
7	m_3	m_1	m_2	m_4

Indeed, w_1 and w_2 are both better off (and, as Theorem 7 states, no other woman is worse off, and even the innocent w_3 is better off as well). However, this lie is clearly not personally optimal for w_2 , since she could declare any order of preference under which her first choice is m_3 , and become even better off by being matched with him. Nonetheless, if she tried to do so (either by telling the truth, or by lying), then w_1 would become worse off, and in this case it would be better for w_1 to tell the truth, which, as stated above, would cause the new matching to be identical to the old matching, and hence would cause w_2 to be matched with m_2 , ending up in a worse situation than had she lied in the above-described non-personally-optimal way.

It should be noted that the resulting new matching is neither stable under the original preferences (as w_2 and m_3 truly prefer each other over their respective matches) nor can it be achieved by simply truncating the preference lists of w_1 and w_2 (as any such truncation will result in w_2 either having a blank list or a list with m_3 as her first choice), giving, as promised, an example in which the proofs from the previous section do not hold.

We conclude the analysis of this example by noting that in the above-described lying game between w_1 and w_2 , the utility for each of the players given these lying strategies is higher than her utility in any Nash equilibrium, and this is the only pair of strategies (up to the degrees of freedom discussed above) with this property. (Moreover, the utility of w_1 is higher in the resulting unstable matching than in any stable matching, since she is matched with m_1 in every stable matching, as can be verified by obtaining the women-optimal stable matching by running the algorithm with the women serenading under the men's windows.)

2.4 A General Proof

Before proving the general case of Theorem 7, let us first introduce some notation: For each person $p \in W \cup M$, let us denote the person from the opposite sex matched with p under the original matching (resp. new matching) by $O(p)$ (resp. $N(p)$).

Lemma 9. *If a man $m \in M$ never serenades under the window of a woman $w \in W$ during NA (resp. OA), then m prefers $N(m)$ (resp. $O(m)$) over w .*

Proof. Since m approaches women according to his order of preference, and since he never reaches w 's window during NA (resp. OA), then he must end up being matched, under the new (resp. old) matching, to a woman he prefers over w , and by definition, that woman is $N(m)$ (resp. $O(m)$). \square

By the following Lemma, we need only prove part (b) of Theorem 7.

Lemma 10. *If a woman $w \in W$ is worse off, then $O(w)$ is better off.*

Proof. Since w is worse off, w prefers $O(w)$ over $N(w)$ according to her true order of preference. Since it is given that no liar is worse off, then w is not a liar, and therefore she declared that she prefers $O(w)$ over $N(w)$ also during NA . Therefore, since w prefers $N(w)$ over any other man who serenades under her window during NA , it follows that $O(w)$ could not have serenaded under her window on any night during NA . Thus, by Lemma 9, $O(w)$ prefers $N(O(w))$ over $w = O(O(w))$, making $O(w)$ better off. \square

We now present a key definition underlying the proof of Theorem 7, and follow with phrasing and proving a sequence of lemmas shedding new light on the structure of lying by women in the Gale-Shapley algorithm and culminating with the proof of Theorem 7.

Definition 11 (Rejecter). A woman $w \in W$ is said to be a *rejecter* if she rejects $N(w)$ during OA . In this case, let the man who serenades under her window on the night of OA on which she rejects $N(w)$, but whom she does not reject on that night (and whom, therefore, she prefers over $N(w)$), be denoted by $B(w)$.

Lemma 12. *If a man $m \in M$ is better off, then $N(m)$ is a rejecter.*

Proof. Since m is better off, he prefers $N(m)$ over $O(m)$. Therefore, since m approaches women according to his order of preference, m would not serenade under $O(m)$'s window before first having been rejected by $N(m)$. Now, since m serenades under $O(m)$'s window during the last night of OA , it follows that he must have been rejected by $N(m)$ on some earlier night during OA . Since $N(m)$ rejects $m = N(N(m))$ during OA , she is a rejecter. \square

Lemma 13. *If a woman $w \in W$ is a rejecter, then she is worse off.*

Proof. By induction and transitivity, w prefers $O(w)$ over any man she rejects during OA and therefore, being a rejecter, over $N(w)$. Thus, w is worse off. \square

Lemma 14. *If a woman $w \in W$ is a rejecter, then $B(w)$ prefers $N(B(w))$ over w .*

Proof. By Lemma 13, w is not a liar, and as such, her order of preference during NA is her true order of preference. Since w is a rejecter, she prefers $B(w)$ over $N(w)$. Therefore, as w truly prefers $N(w)$ over any other man who serenades under her window during NA , it follows that $B(w)$ could not have serenaded under w 's window during NA . Thus, by Lemma 9, $B(w)$ prefers $N(B(w))$ over w . \square

Lemma 15. *If a woman $w \in W$ is a rejecter, then*

1. $N(B(w))$ is a rejecter.
2. During OA , w rejects $N(w)$ on a strictly later night than the night on which $N(B(w))$ rejects $B(w)$.

Proof. Throughout this proof, for the sake of conciseness, wherever we discuss whether a rejection or serenading occurred, or when it occurred, we always refer to OA . Since w is a rejecter, then according to Lemma 14, $B(w)$ prefers $N(B(w))$ over w . Also note that by definition of $B(w)$, $B(w)$ serenades under w 's window on some night. By these two observations, and since $B(w)$ approaches women according to his order of preference, $N(B(w))$ rejects $B(w)$, and does so on a night strictly earlier than any night on which $B(w)$ serenades under w 's window. Since $N(B(w))$ rejects $B(w) = N(N(B(w)))$, then she is a rejecter. Moreover, since, by definition of $B(w)$, w rejects $N(w)$ on a night on which $B(w)$ serenades under her window as well, then this night is a strictly later night than the night on which $N(B(w))$ rejects $B(w)$. \square

To complete the proof of Theorem 7, assume for a contradiction that there exists a better-off man $m \in M$. Let us denote $w_1 = N(m)$, and by Lemma 12, w_1 is a rejecter. Now, for each $i \in \mathbb{N}$, assume by induction that w_i is a rejecter and set $w_{i+1} = N(B(w_i))$. By Lemma 15, w_{i+1} is a rejecter. Moreover, by that lemma, during OA , w_i rejects $N(w_i)$ on a night strictly later than the night on which w_{i+1} rejects $B(w_i) = N(N(B(w_i))) = N(w_{i+1})$.

From finiteness of W , there must exist $i < j$ such that $w_i = w_j$, but by induction, since $i < j$, then during OA , w_i rejected $N(w_i)$ on a night strictly later than the night on which w_j rejected $N(w_j)$ — a contradiction. \square

Corollary 16. *Under the conditions of Theorem 7, every newly-matched couple (w, m) consists of a better-off woman and a worse-off man.*

3 Generalizations

3.1 Polygamous Matchings

In [3], a one-to-many version of the algorithm is specified and proposed as a way to assign students to colleges: On each algorithm-step, the students apply to their favorite college among those that have not yet rejected them, and then each college rejects all applicants except for the most preferred ones, according to its quota. It is proven there that the resulting matching is stable and that it is optimal (within all stable matches) for each applicant. It can also be shown that it is the worst³ (within all stable matches) for each college.

In [11], it is noted that the assignment of medical interns to hospitals in the USA had been performed using a similar algorithm since 1951, however, in this algorithm the roles were switched and the hospitals were the 'proposers'. The resulting matching is thus the worst for each intern and optimal (within all stable matches) for each hospital in a strong sense: if some hospital has quota n , then it is matched with the n interns it prefers most out of the set of all interns matched with it in any stable matching.

In this section, we prove a generalization of Theorem 7 for these scenarios. In order to ease the transition from the previous section, we maintain the notation of women and men, and refer to the above scenarios as the *polygamous scenarios*. We first redefine our notation for these scenarios. Let W and M be finite sets of women and men, respectively, and let each person be endowed, as before, with a strict order of preference with regards to the members of the opposite sex. For each person $p \in W \cup M$, define n_p to be the *quota* of this person, i.e. the number of spouses from the opposite sex this person seeks. As our goal is the generalization of both algorithms described above, the reader may assume, for ease of readability, that either all women are monogamous ($\forall w \in W : n_w = 1$) or all men are monogamous ($\forall m \in M : n_m = 1$), however the rest of this paper holds verbatim even if this is not the case. (If both the women and men are monogamous, we are reduced to the scenario given in the previous section.) We also assume for now that $\sum_{w \in W} n_w = \sum_{m \in M} n_m$. (This guarantees that when the algorithm stops, each person p is matched with exactly n_p people of the opposite sex — this reduces in the monogamous case to W and M being of identical size.)

Definition 17 (Matching (Polygamous Case)). A map between W and M , mapping each $p \in W \cup M$ to exactly n_p members of the set p does not belong to, is called a *matching*.

The definition of instability of a matching requires another detail, which was inferred from the other requirements in the monogamous scenario.

Definition 18 (Stability (Polygamous Case)). A matching is said to be *unstable* under the given orders of preference if there exist two matched couples (w, m) and (\tilde{w}, \tilde{m}) such that:

³See the discussion following Definitions 19 and 20 below for the precise meaning of worst in this context.

1. w is not matched with \tilde{m} .
2. w prefers \tilde{m} over m .
3. \tilde{m} prefers w over \tilde{w} .

As before, a matching that is not unstable is said to be *stable*.

On each night of the polygamous algorithm, each man $m \in M$ serenades under the windows of the n_m women that he prefers most out of all women who have not (yet) rejected him, and then each woman $w \in W$, under whose window more than n_w men serenade, rejects each of these men, except for the n_w men she prefers most among them.

Now, as before, assume that a subset of the women, denoted L , declare false orders of preference for themselves. For a person $p \in W \cup M$, let us denote the set of people of the opposite sex matched with p under the original matching (resp. the new matching) by $O(p) = \{\sigma_1^p, \dots, \sigma_{n_p}^p\}$ (resp. $N(p) = \{n_1^p, \dots, n_{n_p}^p\}$) such that p prefers σ_i^p over σ_{i+1}^p (resp. prefers n_i^p over n_{i+1}^p).

Before we formulate the polygamous version of Theorem 7, we have to redefine the circumstances under which a person is said to be better, or worse, off. It should be noted that while in the monogamous scenario each person's order of preference yields a full order on the set of possible matches for that person (i.e. people of the opposite sex), in the polygamous case each person's order of preference yields a partial order on the set of possible matches for that person (i.e. n_p -tuples of people of the opposite sex). This introduces an asymmetry between the following two definitions, which did not exist in the monogamous scenario.

Definition 19 (Improvement (Polygamous Case)). A person $p \in W \cup M$ is said to be *weakly better off* (resp. *weakly worse off*) if for each $1 \leq i \leq n_p$, p does not prefer σ_i^p over n_i^p (resp. does not prefer n_i^p over σ_i^p).

Definition 20 (Worsening (Polygamous Case)). A person $p \in W \cup M$ is said to have *gained only worse matches* if p prefers every member of $O(p)$ over every member of $N(p) \setminus O(p)$.⁴

We note that by [6, Theorem 1.6.4], for every person p , the relation “ p has gained only worse matches in ... over ...” defines a total order over the equivalence classes of stable matchings (where the set of stable matchings is partitioned into equivalence classes according to the set of spouses to which p is matched). Moreover, that theorem may be used to show that in fact, if we restrict ourselves to stable matchings, having gained only worse matches is equivalent to being weakly worse off, and thus dual to being weakly better off. Nonetheless, as pointed out above, one of the main contributions of our study is in dealing with the case in which the new matching need not be stable. Thus, we do not assume any duality/equivalence between Definitions 19 and 20. Indeed, it is possible to show by means of a simple example that it may be possible for a woman to lie in a manner

⁴Note that this condition is met in the special case in which $N(p) = O(p)$.

that makes her better off (i.e. weakly better off with $N(w) \neq O(w)$) without preferring every member of $N(w)$ over every member of $O(w) \setminus N(w)$ and without preferring every member of $N(w) \setminus O(w)$ over every member of $O(w)$. We now proceed to reformulate Theorem 7 for the polygamous scenario.

Theorem 21 (Sisterhood (Polygamous Case)). *Under the above conditions, if all lying women are weakly better off, then:*

- (a) *All women are weakly better off.*
- (b) *All men have gained only worse matches.*

The special case of Theorem 21, in which all women are polygamous and men are monogamous, can be easily proven by reduction to the conditions of Theorem 7 by “replicating” each polygamous woman $w \in W$ into n_w distinct monogamous women $\{(w, i)\}_{i=1}^{n_w}$, each having the same order of preference as w . For each man $m \in M$, replace w on his list of preferences with these women, in such a way that he prefers (w, i) over $(w, i + 1)$ for all i . A reduction along these lines was used in [2] to generalize certain properties of the monogamous scenario to the polygamous-women scenario and it was shown there that the men matched with $(w, 1), \dots, (w, n_w)$ by the monogamous algorithm are exactly those matched with w by the polygamous algorithm. Furthermore, in the notations of this paper, it can be shown that the man matched with (w, i) by OA is o_i^w , and by NA — is n_i^w . This yields that w is weakly better off if and only if none of the monogamous women (w, i) is worse off. Thus, the reduction is complete, as it is clear that since all men are monogamous, if a man is not better off under the reduction, then he has only gained worse matches before the reduction.

Unfortunately, replicating each man in a similar manner would not yield a proof for even the monogamous women / polygamous men scenario as easily, for it is possible for a woman w to be weakly better off before the reduction by maintaining the same match m , but to be worse off under the reduction because her match is e.g. $(m, 2)$ instead of $(m, 1)$. Also, every replicated monogamous man not being better off under the reduction does not necessarily imply that every polygamous man has gained only worse matches before the reduction. To make things even worse, in the general case where both men and women may be polygamous, running the monogamous algorithm after such a replication does not even produce the same matching as would be produced by the polygamous algorithm, as it may lead to situations such as a couple matched to each other with “multiplicity” greater than 1.

Indeed, in order to prove Theorem 21 in its general form we now retrace our steps and revisit the inner workings of the proof of Theorem 7, rewriting it to allow for the generalizations we introduced. As before, we begin with a lemma establishing that we need only prove part (b) of Theorem 21.

Lemma 22. *If a woman $w \in W$ is not weakly better off, then there exists $m \in O(w)$ who has not gained only worse matches.*

Proof. Since it is given that all liars are weakly better off, it follows that w is not a liar, and therefore her order of preference during NA is her true order of preference. Since w is not weakly better off, there exists $1 \leq i \leq n_w$ such that w prefers o_i^w over n_i^w and thus prefers o_1^w, \dots, o_i^w over n_i^w . By induction and by definition of the polygamous algorithm, there are exactly $i - 1$ men who serenade under w 's window during NA and that she (truly) prefers over n_i^w (these are n_1^w, \dots, n_{i-1}^w), so by the pigeonhole principle, there exists $1 \leq j \leq i$ such that o_j^w does not serenade under w 's window during NA . Therefore, since o_j^w approaches women according to his order of preference, o_j^w prefers all women in $N(o_j^w)$ over w . Since $|N(o_j^w)| = n_{o_j^w} = |O(o_j^w)|$, and since $w \in O(o_j^w) \setminus N(o_j^w)$, it follows that there exists $\tilde{w} \in N(o_j^w) \setminus O(o_j^w)$ and as stated, o_j^w prefers her over w , and hence (as $w \in O(o_j^w)$) o_j^w has not gained only worse matches. \square

Before continuing, we now refine the definition of a rejecter for the polygamous scenario.

Definition 23 (Rejecter (Polygamous Case)).

1. A woman $w \in W$ is said to be a *rejecter* if she rejects any of the members of $N(w)$ during OA . We denote the set of all such rejected members of $N(w)$ by $R(w)$.
2. A man $m \in M$ is said to be a *rejectee* if there exists a rejecter $w \in N(m)$ such that $m \in R(w)$.

Similarly to the proof of Theorem 7, a key role in the proof of Theorem 21 is played by a man $B(w, m)$ for whom, in a sense, w rejected m . We defer the precise definition of this man to Lemma 26.

Lemma 24. *If a man $m \in M$ has not gained only worse matches, then he is a rejectee.*

Proof. Since m has not gained only worse matches, there exist $w \in O(m)$ and $\tilde{w} \in N(m) \setminus O(m)$ such that m prefers \tilde{w} over w . Since $w \in O(m)$, it follows that m serenades under w 's window during OA . Since m approaches women according to his order of preference, he would never serenade under w 's window during OA without having serenaded on the same night, or on a previous night, under \tilde{w} 's window. However, since m is not matched with \tilde{w} at the end of OA , it must hold that \tilde{w} rejects him during OA and therefore, by definition, $m \in R(\tilde{w})$, and thus m is a rejectee. \square

Lemma 25. *If a woman $w \in W$ is a rejecter, then she is not weakly better off.*

Proof. Since w is a rejecter, there exists $r \in R(w)$. By definition of $R(w)$, there exists $1 \leq i \leq n_w$ such that $n_i^w = r$. Since, by definition of $R(w)$, w rejects r during OA , and since, by induction and transitivity, w prefers each member of $O(w)$ over each man she rejects during OA , it follows that she prefers o_i^w over $r = n_i^w$, and is, thus, not weakly better off. \square

Lemma 26. *If a woman $w \in W$ is a rejecter, then for each $r \in R(w)$ there exists a man $B(w, r)$ such that*

1. $B(w, r)$ serenades under w 's window during OA on the night on which she rejects r , but $B(w, r)$ is not rejected by her on that night.
2. $B(w, r)$ prefers each member of $N(B(w, r))$ over w .

(If more than one such man exists, define $B(w, r)$ to be one of these men, arbitrarily.)

Proof. By Lemma 25, w is not a liar, and as such, her order of preference during NA is her true order of preference. Let B be the set of all men who serenade under her window on the night during OA on which she rejects r , but who were not rejected by her on that night. By definition of the Gale-Shapley algorithm, w prefers each member of B over r , and $|B| = n_w$. Since w does not reject r during NA (since $r \in N(w)$), and since the order of preference of w during NA is her true order of preference, it follows that not all members of B serenade under her window during NA (for she is matched under NA to the set of the n_w men that she prefers most out of all the men who serenade under her window during NA). Define, therefore, $B(w, r)$ to be a member of B who does not serenade under w 's window during NA . Since $B(w, r)$ does not serenade under w 's window during NA , and since he approaches women according to his order of preference, it follows that he prefers each member of $N(B(w, r))$ over w . \square

Lemma 27. *If a woman $w \in W$ is a rejecter, then for each $r \in R(w)$ there exists a woman $\tilde{w} \in N(B(w, r))$ such that*

1. \tilde{w} is a rejecter.
2. $B(w, r) \in R(\tilde{w})$
3. During OA , w rejects r on a strictly later night than the night on which \tilde{w} rejects $B(w, r)$.

Proof. Once again, throughout this proof, for the sake of conciseness, wherever we discuss whether a rejection or serenading occurs, or when it occurs, we always refer to OA . Since w is a rejecter, then by Lemma 26, $B(w, r)$ serenades under w 's window (on the night on which she rejects r) and $B(w, r)$ prefers each member of $N(B(w, r))$ over w . Hence, as $B(w, r)$ approaches $n_{B(w, r)}$ women each night according to his order of preference, and as $|N(B(w, r))| = n_{B(w, r)}$, it follows that $B(w, r)$ serenades on earlier nights under each of the windows of $N(B(w, r))$ and is rejected by at least one of them on a night strictly earlier than any night on which he serenades under w 's window — let us denote such a woman by \tilde{w} . Since \tilde{w} rejects $B(w, r)$ (with whom she is matched under the new matching), it follows that she is a rejecter and that $B(w, r) \in R(\tilde{w})$. Moreover, since, by definition of $B(w, r)$, w rejects r on a night on which $B(w, r)$ serenades under her window as well, then this night is a strictly later night than the night on which \tilde{w} rejects $B(w, r)$. \square

To complete the proof of Theorem 21, assume for a contradiction that there exists a man $m_1 \in M$ who has not gained only worse matches. By Lemma 24, m_1 is a rejectee, therefore there exists a rejecter $w_1 \in N(m_1)$ such that $m_1 \in R(w_1)$. Now, for each $i \in \mathbb{N}$,

assume by induction that w_i is a rejecter and that $m_i \in R(w_i)$ and set, by Lemma 26, $m_{i+1} = B(w_i, m_i)$. By Lemma 27, there exists a rejecter $w_{i+1} \in N(m_{i+1})$. Moreover, by that lemma, during OA , w_i rejects m_i on a night strictly later than the night on which w_{i+1} rejects m_{i+1} .

From finiteness of $W \times M$, there must exist $i < j$ such that $w_i = w_j$ and $m_i = m_j$, but by induction, since $i < j$, then during OA , w_i rejects m_i on a night strictly later than the night on which w_j rejects m_j — a contradiction. \square

3.2 Blacklists and Mismatched Quotas

As mentioned before, in [3] it is not required that the sum of the quotas of all colleges be the same as the number of applicants, resulting in some colleges not fulfilling their quotas or some applicants not being accepted to any college. (In the monogamous scenario, for instance, this translates to W and M not necessarily being of equal size, a variant first explicitly studied in [9].) Moreover, it is allowed for a college to remove some of the students from its preference list, indicating that the college is unwilling to accept these candidates even if it means that its quota is not met. Similarly, applicants are allowed to remove some of the colleges from their preference lists, indicating their unwillingness to attend these colleges even at the risk of not being accepted to any college.

The modified algorithm-step for this scenario (in the notations used throughout this document) is that if, by a certain night, the number of women who have not blacklisted or (yet) rejected a man $m \in M$, and are not blacklisted by him, is less than n_m , then on that night he serenades under the windows of all these women. (Otherwise, if there are at least n_m such women, then he serenades, as before, under the windows of the n_m women that he prefers most out of them.)

Let us adjust our notations for this scenario, and then prove that the appropriate generalization of Theorems 7 and 21 still holds under it.

Definition 28 (Matching (General Case)). A map between W and M , mapping each $p \in W \cup M$ to **at most** n_p members of the set p does not belong to, is called a *matching*.

Definition 29 (Stability (General Case)). A matching is said to be *unstable* under the given orders of preference if either of the following hold.

- The conditions of Definition 18 are met.
- There exists a matched couple (w, m) and a woman (resp. man) p such that all of the following hold.
 1. p is matched with less than n_p spouses.
 2. p has not blacklisted m (resp. w).
 3. m (resp. w) prefers p over w (resp. m).
- There exist a woman w and a man m , neither of whose quotas are filled, and neither of whom has blacklisted the other.

Once again, a matching that is not unstable is said to be *stable*.

For a person $p \in W \cup M$, let us define $O(p)$ and $N(p)$ as before, and for each $1 \leq i \leq |O(p)|$ (resp. $1 \leq i \leq |N(p)|$), let us define o_i^p (res. n_i^p) as before as well.

Definition 30 (Improvement (General Case)). A woman $w \in W$ is said to be *weakly better off* if the following conditions hold.

1. $N(w)$ contains none of the men blacklisted by w .
2. $|O(w)| \leq |N(w)|$
3. For each $1 \leq i \leq O(w)$, w does not prefer o_i^w over n_i^w .

If, in addition, either for some $1 \leq i \leq O(w)$, w prefers n_i^w over o_i^w (w has *improved her matches*), or $|O(w)| < |N(w)|$ (w has *gained matches*), then w is said to be *better off*.

The definition of a man having gained only worse matches remains unchanged. Specifically, it should be emphasized that this definition does not require that $|N(m)| \leq |O(m)|$. Nonetheless, we show in Corollary 32 that under the above conditions the inverse is an impossibility. Similarly, the same corollary shows that under the above conditions, a woman may not be better off due to gaining matches, but only due to improving matches. It should also be noted that, by definition of the Gale-Shapley algorithm, it is not possible for any person declaring their true preferences to be matched, under the new matching, to any person blacklisted by them. As we now show, a generalized version of Theorems 7 and 21 for this general scenario can be proven by way of reduction to Theorem 21.

Corollary 31 (Sisterhood (General Case)). *Theorem 21 holds under the above conditions.*

Proof. We prove Corollary 31 by reducing to the conditions of Theorem 21 by introducing, for each person p , n_p new monogamous people of the opposite sex $\emptyset_1^p, \dots, \emptyset_{n_p}^p$, each of whom prefers p the most (the rest of their orders of preference can be arbitrary). The original (resp. new) order of preference for p will now be: first, all of the original people of the opposite sex who are not originally (resp. newly) blacklisted by p , ordered according to her or his given original (resp. new) order of preference, then $\emptyset_1^p, \dots, \emptyset_{n_p}^p$ in this order, and then, in an arbitrary order, the original (resp. new) blacklist of p and the people newly-introduced for the other people of the same sex as p . It is straight-forward to verify that the quotas of all women (original and newly-introduced) match the quotas of all men (original and newly-introduced). It is left for the reader to verify that the old (resp. new) matching before the reduction is identical to the old (resp. new) matching under the reduction, after the newly-introduced people have been removed from it, that each of the original women is weakly better off before the reduction iff she is weakly better off under the reduction, and that if an original man has gained only worse matches under the reduction, then he has also gained only worse matches before the reduction. \square

Corollary 32. *Under the conditions of Corollary 31, $|N(p)| = |O(p)|$ for each person $p \in W \cup M$.*

Proof. Observe that for each person $p \in W \cup M$, their old (resp. new) matches under the reduction are their old (resp. new) matches before the reduction, “padded” by as many people as needed from the top of $\emptyset_1^p, \dots, \emptyset_{n_p}^p$ to fulfill their quota of n_p matches. Assume, for a contradiction, that there exists a man $m \in M$ such that (before the reduction) $|N(m)| > |O(m)|$. Let $o = |O(m)|$ (before the reduction), then under the reduction, m is matched with $\emptyset_{n_p-o}^p$ under the original matching, but not under the new matching. Recall that $\emptyset_{n_p-o}^p$ prefers p the most, so she is not weakly better off — a contradiction. We have thus shown that for each man $m \in M$, $|N(m)| \leq |O(m)|$. Since all women are weakly better off, for each woman $w \in W$ we have $|N(w)| \geq |O(w)|$. Thus, by both of these, we have

$$\sum_{w \in W} N(w) \geq \sum_{w \in W} O(w) = \sum_{m \in M} O(m) \geq \sum_{m \in M} N(m) = \sum_{w \in W} N(w),$$

which yields that all expressions in this sequence of inequalities are actually equal. Since for every woman $w \in W$, $|N(w)| \geq |O(w)|$, (resp. for every man $m \in M$, $|N(m)| \leq |O(m)|$), and since summing over all women (resp. all men) we get an equality, it follows that the equality holds for each woman (resp. each man) separately. \square

It should be noted that in [5], Corollary 32 is proved using an entirely different approach for the special case of the monogamous scenario in which the new matching is stable, and in [12], it is proved also for the polygamous scenario where the new matching is stable. Corollary 32 generalizes these results, as it does not require stability of the new matching.

Corollary 33. *Under the conditions of Theorem 7, if $|L| = 1$ and the lying woman is better off, then so is some innocent woman too.*

Proof. Since the liar is better off, she is newly matched with some man to whom she was not originally matched, and since the number of matches for this man remains the same, he is no longer matched with some woman to whom he was originally matched, and hence the set of matches for that woman has changed, and thus, since she is weakly better off, she is better off. \square

4 Summary and Open Problems

Throughout this paper, we have shown that even in very general scenarios, a form of sisterhood exists in the Gale-Shapley matching algorithm, both in terms of women not harming each other, and in terms of them not helping any man.

Examining Definitions 19 and 20, one may find that having gained only worse matches is stronger than (i.e. a special case of) being weakly worse off (which, in turn, is the dual of being weakly better off). In other words, in the scenarios discussed in this paper, in a sense, the minimum possible damage to a man is greater than the minimum possible gain by a woman. It would be interesting to establish whether, under various conditions, the overall damage to men is, in any sense, usually greater than the overall gain for women

(and thus resulting in damage to the entire population as a whole, which contrasts with the sisterhood that exists amongst the women).

The example given in Section 2.3 (as well as other examples given in the first author's undergraduate thesis (2007), upon which this paper is based) had to be crafted very delicately. It would be interesting to establish whether, when the orders of preference are determined by a random model, lies can be very beneficial (either for the lying women, or for their innocent colleagues), or whether the utility gained from such a lie is usually relatively small.⁵ Furthermore, in many of these examples, and in many examples from the literature, the preferences of every person differ greatly from those of each of their colleagues. In the real-world scenario of colleges and applicants, however, it is reasonable to expect the preferences of many applicants to be similar, and the same applies to the preferences of many colleges. There will always be differences, but it seems that they are likely to be local, such as a permutation on colleges that are adjacent to each other in the order of preference. It seems unlikely, for example, for a certain college to be highly rated by half of the applicants and poorly rated by the rest. (All this applies to the preferences of colleges as well, naturally.) It would be interesting to try and build a probabilistic model along these lines, that matches the observed behaviors by medical interns and colleges throughout the years, and to check the questions raised above under such a model. In addition, since in the real world the information each player has is not perfect, it could be interesting to check whether, given only knowledge of a statistical model for the preferences of the rest of the players (and possibly full knowledge of the preferences of just a few players), there exist strategies (for individual players, or for small sets of players) that are expected to be better for these players than telling the truth. It could also be interesting to check what happens if such strategies are simultaneously applied by more than one conspiring set of players.

Acknowledgements

We would like to thank Sergiu Hart for useful conversations and for providing us with references to the literature. We would like to thank the anonymous referee for references to the literature as well as for careful reading and many helpful remarks regarding the presentation.

⁵Some work has already been done in this area: In [8], it is shown that if the preference lists of all men are of a small size, independent of $|W|$ and $|M|$, then the expected number of people with more than one stable spouse is vanishingly small (as $|W|$ and $|M|$ approach infinity), regardless of the distribution of the preference lists; in [13], it is shown by simulation that under a certain symmetry condition between the preference lists of the women and the men, the number of women who can benefit from lying with no accomplices is small.

References

- [1] G. Demange, D. Gale and M. Sotomayor. A further note on the stable matching problem. *Discrete Applied Mathematics*, 16(3):217–222, 1987.
- [2] L. E. Dubins and D. Freedman. Machiavelli and the Gale-Shapley algorithm. *American Mathematical Monthly*, 88(7):485–494, 1981.
- [3] D. Gale and L. S. Shapley. College admissions and the stability of marriage. *American Mathematical Monthly*, 69(1):9–15, 1962.
- [4] D. Gale and M. Sotomayor. Ms. Machiavelli and the stable matching problem. *American Mathematical Monthly*, 92(4):261–268, 1985.
- [5] D. Gale and M. Sotomayor. Some remarks on the stable matching problem. *Discrete Applied Mathematics*, 11(3):223–232, 1985.
- [6] D. Gusfield and R. W. Irving. *The Stable Marriage Problem: Structure and Algorithms*. MIT Press, 1989.
- [7] C.-C. Huang. Cheating by men in the Gale-Shapley stable matching algorithm. In *Proceedings of the 14th Annual European Symposium on Algorithms (ESA)*, pages 418–341, 2006.
- [8] N. Immorlica and M. Mahdian. Marriage, honesty, and stability. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 53–62, 2005.
- [9] D. G. McVitie and L. B. Wilson. Stable marriage assignment for unequal sets. *BIT* 10(3):295–309, 1970.
- [10] D. G. McVitie and L. B. Wilson. The stable marriage problem. *Communications of the ACM* 14(7):486–490, 1971.
- [11] A. E. Roth. The evolution of the labor market for medical interns and residents: a case study in game theory. *Journal of Political Economy* 92(6):991–1016, 1984.
- [12] A. E. Roth and M. A. O. Sotomayor. *Two-Sided Matchings*. Cambridge University Press, 1990.
- [13] C.-P. Teo, J. Sethuraman and W.-P. Tan. Gale-Shapley stable marriage problem revisited: strategic issues and applications. *Management Sciences* 47(9):1252–1267, 2001.