Fair Stable Matching Meets Correlated Preferences

Angelina Brilliantova
Rochester Institute of Technology
Rochester, NY, USA
1b9849@rit.edu

Hadi Hosseini
Pennsylvania State University
University Park, PA, USA
hadi@psu.edu

Abstract

The stable matching problem sets the economic foundation of several practical applications ranging from school choice and medical residency to ridesharing and refugee placement. It is concerned with finding a matching between two disjoint sets of agents wherein no pair of agents prefer each other to their matched partners. The Deferred Acceptance (DA) algorithm is an elegant procedure that guarantees a stable matching for any input; however, its outcome may be unfair as it always favors one side by returning a matching that is optimal for one side (say men) and pessimal for the other side (say women). A desirable fairness notion is minimizing the sex-equality cost, i.e. the difference between the total rankings of both sides. Computing such stable matchings is a strongly NP-hard problem, which raises the question of what tractable algorithms to adopt in practice. We conduct a series of empirical evaluations on the properties of sex-equal stable matchings when preferences of agents on both sides are correlated. Our empirical results suggest that under correlated preferences, the DA algorithm returns stable matchings with low sex-equality cost, which further confirms its broad use in many practical applications.

1 Introduction

Matching theory sets the economic foundation for achieving stable allocations through market design. It has shaped the cornerstone of many practical applications ranging from school choice [2, 4] and medical residency [62, 64] to refugee placement [19, 38] and ridesharing [9, 28]. In its essence, the stable matching problem deals with finding a matching between two disjoint sets of agents (colloquially men and women) according to their preferences. The primary objective is to achieve stability, that is, finding a matching between the two sides wherein no pair of agents prefer each other to their matched partners.

Over the past few decades, numerous theoretical breakthroughs and developments were pioneered to study mathematical and axiomatic properties of the stable matching problem [20, 24, 61, 69] as well as its computational and algorithmic aspects [30, 40, 45, 67, 72]. However as problems become increasingly more challenging, there has been a need for moving from theory about simple markets to more complex settings that account for subtle, but vital, differences in constraints in preference structures or other factors that may pose computational or axiomatic challenges. Hence, as Roth and Peranson [64] suggested “as game theory moves from simple conceptual problems to complex design problems, we will need to make more general use of this interaction among theory, computational investigation of market data, and theoretical computation, and that this in turn will produce new problems and directions for traditional theory”. This doctrine motivated a large body of work in taking empirical or statistical approaches in exploring markets through studying (real or synthetic) data sets or analyzing statistical distributions [6, 13, 44, 71].

In this vein, we investigate the fairness of stable matchings through empirical simulations to paint a thorough picture of the structure of fair stable solutions in matching markets.
A motivating insight

The Deferred Acceptance algorithm (DA)—due to Gale and Shapley [23]—provides an elegant solution to the stable matching problem wherein agents from one side (say men) make proposals to the agents from the other side (say women). Each woman tentatively accepts her favorite proposal and rejects the rest. Despite the popularity and success of the DA algorithm in many real-world matching markets [61, 64], it always favors the proposing side to the receiving side, that is, the DA algorithm always returns a stable matching that is men-optimal [23] but women-pessimal [54]. This unequal treatment of the sides raises critical questions about the fairness of the DA algorithm, which may result in (extremely) unequal welfare between both sides.

One of the most prominent and well-studied fairness notions—proposed by Gusfield and Irving [27]—is sex-equality that aims at finding a matching that equalizes the welfare of both sides by minimizing the difference between the total rankings of men and women in a stable matching. While sex-equal stable matchings always exist, computing one has shown to be strongly NP-hard [40]. For any instance of the matching problem, there may be an exponential number of stable solutions [35] and these stable matchings form a distributive lattice. Thus, one may hope to find a “fair” stable matching that equalizes the welfare of both sides.

On a closer scrutiny, however, we notice that a sex-equal stable matching is highly correlated with the structure of preference lists of each side of the market. This observation suggests that even though the number of stable matchings grows exponentially in general [35], surprisingly a sex-equal solution lies at the extreme points of the stable lattice in certain settings.

Example 1.1. Consider the following instance with five men and five women and the following preference lists.

\[
\begin{align*}
m_1 &: w_2^\dagger \ w_4^* \ w_5 \ w_1 \ w_3 \ w_1^* \ m_4 \ m_2 \ m_1 \ m_5^* \ m_3^\dagger \\
m_2 &: w_2^\dagger \ w_4 \ w_2 \ w_1 \ w_5 \ w_2^* \ m_4^* \ m_4 \ m_1^\dagger \ m_5 \ m_3 \\
m_3 &: w_1^\dagger \ w_5 \ w_4 \ w_3 \ w_2 \ w_3^* \ m_4^* \ m_2^\dagger \ m_1 \ m_3 \ m_5 \\
m_4 &: w_4^\dagger \ w_2 \ w_3 \ w_1 \ w_5 \ w_4 \ m_2 \ m_1^\dagger \ m_4^\dagger \ m_5 \ m_3 \\
m_5 &: w_2 \ w_3 \ w_5^\dagger \ w_1^* \ w_4 \ w_5 \ m_1 \ m_4 \ m_2 \ m_3^* \ m_5^\dagger \\
\end{align*}
\]

The men-optimal and women-optimal matchings are marked by † and * respectively. The stable lattice of this preference profile contains five matchings as illustrated in Figure 1. A sex-equal stable matching
assigns a total ranking of 12 to women and 13 to men, minimizing the welfare difference between the two
sides. Notice that the sex-equal solution for this profile lies precisely on the extreme point of the stable
lattice and is equivalent to the women-optimal solution.\(^1\)

Motivated by this observation, we study stable matching problems when both sides of the market have
correlated preference lists. We focus on correlated random preference lists sampled from Mallows distribu-
tion models [50, 51]. The Mallows model is the cornerstone of a variety of ranking problems in machine
learning [8, 49, 75] and social choice [34], and has been shown to correctly capture the preference rank-
ings of individuals in many practical applications [15, 68]. Under the Mallows models, preferences are
correlated through a reference ranking, where the probability of a preference list to be sampled is inversely
proportional to its Kendall-tau distance [41, 42] (number of swaps between the two lists) from the reference
ranking. Therefore, focusing on settings with correlated preference lists on both sides of the market we ask
the following questions:

How does the correlation between preference lists of two sides impact the fairness of stable
matchings? What algorithmic solutions should we adopt in practice when preferences are cor-
related according to Mallows models?

1.1 Our Results

Focusing on markets with correlated preferences on both sides, we empirically investigate the sex-equality
of stable matchings through a series of extensive experiments on synthetic data:

- **Stable lattice**: We first focus on the size of the stable lattice when preferences of both sides are drawn
  from Mallows distributions. We show that the size of the stable lattice is heavily dependent on the
  relationship between the correlation intensity in preferences of two sides (Section 3). In particular,
  we show that when correlations between the two sides of the market are symmetric, that is, prefer-
  ences are distributed according to the same dispersion parameter, the stable lattice grows rapidly—as
  it was theoretically proved by Levy [48]. However, when correlations are asymmetric, i.e. distributed
  according to different dispersion parameters, the size of the stable lattice sharply decreases.

- **Asymmetric correlations**: When preferences of the two sides are sampled from different Mallows
distributions, we show that in overwhelming majority of cases, a sex-equal stable matching is located
at the extreme points of the stable lattice. In Section 4.1 we discuss how this key observation im-
nediately results in a polynomial-time algorithm for finding a sex-equal matching by computing a
men-optimal or women-optimal stable matching through the DA algorithm [23].

- **Symmetric correlations**: When preferences are sampled from the same Mallows distribution, even
  though the size of the stable lattice may be exponential, the cost of a sex-equal matching is consider-
  ably close to the cost of the DA outcome. As the difference between the welfare of men and women
  is sufficiently small, the importance of which group proposes (men or women) becomes negligible
  (Section 4.2).

- **Comparing the performance of algorithms**: In Section 5 we conduct a series of empirical compar-
  isons between well-studied heuristic and procedurally fair algorithms. We show that with correlated
  preferences, the DA algorithm—with a careful selection of the proposing side—performs as good as

\(^1\)We refer the readers to Appendix A for details of lattices and rotations.
the best known local search algorithm with respect to the sex-equality cost even on large instances. This result further justifies the use of the DA algorithm which is more computationally efficient compared to other methods.

1.2 Related Work

Stability is a key condition for the success and longevity of two-sided markets [63]. The seminal work by Gale and Shapley showed that a stable matching always exists and can be found in polynomial time using the Deferred Acceptance algorithm (DA) [23]. The important property of DA is inherent asymmetry: favouring one side at the cost of another [27]. The inherent bias of DA should be taken into account as many central clearinghouses use it as the basis for their matching procedures, including the National Resident Matching Program (NRMP) [64] and the New York City school assignment [3].

Several papers investigate the difference between the expected rank of partners in an optimal and a pessimal stable matching (aka the welfare gap) [6, 7, 58, 59]. In one-to-one balanced markets (i.e. with an equal number of men and women) and when preferences drawn uniformly at random, the expected total rank of agents asymptotically approaches \( n \log n \) in their optimal matching and \( n^2 / \log n \) in their pessimal matching [58]. The same order of scores is preserved in markets with constant tier-based preferences, in which agents rank partners proportionally to their real-numbered popularity scores [7].

Large real-world markets usually admit an exceedingly small number of stable matchings. For example, for the NRMP hospital-resident matching and for Boston Public School student admission there were only a couple of stable matchings each year [57, 64]. It was shown theoretically, that a lattice becomes essentially a singleton in large one-to-one markets, in which one side draws shorter fixed-sized preference lists from an arbitrary distribution [32]. Similar results were demonstrated for many-to-one settings: balanced [46] and unbalanced [6]. At the same time, large stable lattices were theoretically proven in markets with correlated preferences induced by the Watts–Strogatz “small world” model (even when unbalanced) [59], and the Mallows model [48]. Empirically, large lattices are found in settings of matching-with-contracts [29].

Fairness and stability.

The potential exponential size of stable solutions [35] alongside the inherent bias of DA, has motivated the discussion of fairness in stable matching markets. Various algorithms were proposed to ensure fairness towards individuals [21, 22], socio-economic groups [1, 31, 56], regions [39], and sides (e.g. hospitals vs. residents) [53, 60, 70]. Below, we will briefly discuss a variety of fairness notions that have received attention in the literature of stable matchings.

The egalitarian stable matching, is a stable matching that minimizes the sum of rankings of all partners for men and women, hence maximizes the total welfare of all agents [36]. The minimum regret stable matching minimizes the highest rank of agents, therefore minimizing the rank of a partner for the most unsatisfied agent. Egalitarian and minimum regret stable matchings can be found in polynomial time [21, 36], but do not ensure cross-sided fairness: they both can substantially favor one of the sides, by optimizing the welfare on the level of the whole system (egalitarian), and individual agents (minimum regret). In the median stable matching, each agent is matched to its middle favorite partner across all partners from stable matchings; the resulting matching is a median element of a stable lattice. Finding a median stable matching is NP-hard in general, but it could be done in polynomial time in some restricted families of the rotation poset [17]. The median stable matching might not satisfy cross-sided fairness if one side prefers their middle stable partners substantially more than the other side. A procedurally fair algorithm aims to provide agents an equal probability to affect the resultant stable matching by issuing proposals [71], reducing the set of
stable matchings [5], or satisfying blocking pairs [65]. Procedural fairness does not necessarily result in cross-sided fair matchings, and thus, one side may receive much higher welfare compare to the other side [43].

A sex-equal stable matching minimizes the gap between the sum of partners’ ranks of men and women. The sex-equal stable matching problem is an NP-hard problem and fixed parameter tractable with respect to the treewidth of the Hasse diagram of the rotation poset [25] and \( k \) parameter in the \( k \)-range model [18]. To tackle the sex-equal stable matching problem various heuristics have been proposed: for example, performing local search series on a stable lattice (iBILS) [74], and transforming a stable matching using genetic algorithm [55]. While some of these approximate algorithms (e.g. iBILS) perform well in experiments, they bear no theoretical bounds on the quality of an outcome.

Correlated and random preferences.

Several recent works in stable matching have focused on investigating a variety of natural structures of profiles, from profiles generated uniformly at random [6, 58] to random profiles with soft and hard constraints on rankings [7, 44, 71]. The Mallows distribution model drew the attention of researchers as it was shown to realistically capture the preference of agents in several applications involving individual decision-makers [68]. In the Mallows model, preferences are correlated through a reference ranking, representing the objective order of agents’ attractiveness. In contrast to the tiered and random utility models, the Mallows model generates a variety of correlated preferences using only a few parameters: two dispersion parameters and two reference rankings; one of each for each side of the market.

2 Preliminaries

We start by providing a formal representation of the model and define the necessary properties.

Problem setup. An instance of the stable matching problem is specified by the tuple \( I = \langle M, W, \succ \rangle \), where \( M \) is a set of \( n \) men, \( W \) is a set of \( n \) women, and \( \succ \) is a preference profile which consists of the preference lists of all men and women. The preference list of any man \( m \in M \), denoted by \( \succ_m \), is a strict total order over all women in \( W \) (for any woman \( w \in W \), the list \( \succ_w \) is defined analogously). The rank of woman \( w \) in man \( m \)'s preference list, \( \succ_m \), is denoted by \( r(w,m) \). For instance, given \( \succ = (w_2, w_1, w_3) \), we say \( w_2 \) is ranked first in \( m \)'s preference list, i.e., \( r(w_2, m) = 1 \). Similarly, \( r(m, w) \) denotes the rank of \( m \) in \( \succ_w \).

Stable matchings. A perfect matching is a mapping \( \mu : M \cup W \to M \cup W \) such that \( \mu(m) \in W \) for all \( m \in M \), \( \mu(w) \in M \) for all \( w \in W \), and \( \mu(m) = w \) if and only if \( \mu(w) = m \). Given a matching \( \mu \), a man-woman pair \( (m, w) \) is called a blocking pair, with respect to the preference profile \( \succ \), if they prefer each other to their assigned partners under \( \mu \), i.e., \( w \succ_m \mu(m) \) and \( m \succ_w \mu(w) \). A matching is stable if it contains no blocking pairs.

Given a preference profile \( \succ \), the set of all corresponding stable matchings, \( S_{\succ} \), forms a distributive lattice.\(^2\) The maximum and minimum points of a stable lattice correspond to the men-optimal (\( \mu^M \)) and the women-optimal (\( \mu^W \)) matchings respectively [27], i.e. a matching where all men (respectively women) receive their best stable partners. A men-optimal matching is simultaneously women-pessimal: it matches

\(^2\)See Appendix A for a detailed explanation of the distributive lattice and an example.
all woman to their worst stable partners [54]. Figure 1 illustrates a lattice consisting of six stable matchings. The expected size of a stable lattice (correspondingly the size of \( S_{\succ} \)) is asymptotic to \( e^{-1} \ln n \) when preferences are drawn uniformly at random [58] and may grow exponentially with the number of agents [35].

**Welfare measure.** Let \( S_M(\mu) \) denote the sum of ranks of men’s partners in matching \( \mu \), that is, \( S_M(\mu) = \sum_{m \in M} r(\mu(m), m) \). Similarly for women, we let \( S_W(\mu) \) be the sum of ranks of women’s partners given matching \( \mu \), i.e. \( S_W(\mu) = \sum_{w \in W} r(\mu(w), w) \). Note that the smaller values indicate higher social welfare. Thus, a men-optimal matching \( \mu^M \) is the one that minimizes the sum of rankings for men, that is, \( \mu^M = \arg \min_{\mu \in S_{\succ}} S_M(\mu) \). And similarly, for the women-optimal matching we have \( \mu^W = \arg \min_{\mu \in S_{\succ}} S_W(\mu) \). For simplicity, we will refer to the scores of a men-optimal matching \( S_M(\mu^M) \) as \( S_M \)-optimal and \( S_W \)-pessimal scores. Similarly, for a women-optimal matching we write \( S_M \)-pessimal and \( S_W \)-optimal to indicate \( S_M(\mu^W) \) and \( S_W(\mu^W) \).

**The Deferred Acceptance algorithm.** Given a preference profile \( \succ \), the Deferred Acceptance (DA) algorithm, proposed by Gale and Shapley [23], consists of rounds of proposal and rejection phases and proceeds as follows: In each round every man who is currently unmatched proposes to his favorite woman from among those who have not rejected him yet. Each woman tentatively accepts her favorite proposal and rejects the rest. The algorithm terminates in \( O(n^2) \) when no further proposals can be made.

Given any profile \( \succ \) as input, the DA algorithm is guaranteed to return a stable matching [23]. Moreover, the DA algorithm is simultaneously men-optimal [23] and women-pessimal [54], i.e. men receive their favorite stable partners among all matchings available to them in \( S_{\succ} \), and women receive their least favorite stable partners. The DA algorithm returns the stable matchings at the extreme points of the stable lattice depending on which set (men or women) are proposers. Thus, the DA algorithm with women proposing is respectively women-optimal and men-pessimal.

### 2.1 Fair Stable Matchings

Given a matching \( \mu \), the sex-equality cost of \( \mu \) is the absolute difference between the total welfare of men and women, i.e. \( S_M \) and \( S_W \) scores. Formally, the sex-equality cost of a matching \( \mu \) is defined as

\[
c(\mu) = |S_M(\mu) - S_W(\mu)|.
\]

(1)

Given preference profile \( \succ \), a sex-equal stable matching is a matching in \( S_{\succ} \) that minimizes the sex-equality cost across the stable lattice, that is,

\[
\mu^* \leftarrow \arg \min_{\mu \in S_{\succ}} (c(\mu)).
\]

(2)

Intuitively, in a sex-equal stable matching the total rank of men’s partners is as close as possible to that of women (subject to the stability condition). Kato [40] showed that for an arbitrary preference profile the problem of finding a sex-equal stable matching is strongly NP-hard. However, it is fixed parameter tractable with respect to the range of the profile – a metric showing the maximum discrepancy between an agent’s worst and best rank in the preference profile [18]. Also, it can be found efficiently when agents have a specific two-dimensional single-peaked model [66] or if preferences are identical on one side [37].
2.2 Correlated Preferences

There are several plausible ways to study preference models generated from uniform distributions [6, 58] or correlated preferences [7, 10, 59, 71]. The vast majority of these works focused on profiles wherein the preferences of one side are correlated while the other side is considered uniform. We focus on a more general distribution models where both sides of the market have correlated preferences through the Mallows model.

The Mallows model A Mallows distribution is a distance-based probabilistic model for permutations correlated with some common reference [50]. It is parameterized by a reference ranking, \( \hat{\pi} \), and a dispersion parameter \( \phi \in (0, 1] \). Let \( \pi \) be a permutation of a preference list. For any permutation \( \pi \), the Mallows model specifies a probability as follows:

\[
p(\pi | \hat{\pi}, \phi) = \frac{1}{Z} \phi^{\tau(\pi, \hat{\pi})}
\]

where \( \tau(\pi, \hat{\pi}) \) is a Kendall-tau distance (the number of pairwise inversions) between \( \pi \) and \( \hat{\pi} \) and \( Z \) is a normalization constant with \( Z = 1(1 + \phi)(1 + \phi + \phi^2) \ldots (1 + \phi + \ldots + \phi^{n-1}) \).

Note that the dispersion parameter \( \phi \) indicates the ‘intensity’ of correlation between the sampled preferences. When \( \phi = 0 \), the correlation is maximal, that is, the distribution mass is entirely on the reference ranking and all preference lists are identical. When \( \phi = 1 \), the correlation is minimal, the probability mass is distributed evenly between all possible permutations and the Mallows model is equivalent to the Uniform distribution model (also known as Impartial Culture [12] in the computational social choice literature).

For each set involved in the stable matching problem, we consider an independent probabilistic preference model: one Mallows model for the set of men parameterized by \( \hat{\pi}_m \) and \( \phi_m \) and another preference model for women specified by \( \hat{\pi}_w \) and \( \phi_w \). In this way, the preferences of men (similarly women) are globally correlated with \( \hat{\pi}_m (\hat{\pi}_w) \).\(^3\)

Symmetric and asymmetric models. We let \( \phi_\Delta = |\phi_m - \phi_w| \) to represent the correlation disparity in preferences of men and women sampled for this particular instance. Simply put, correlation disparity reflects how similar the probabilistic preferences of men are in comparison to the similarity of women’s preferences. We call stable matching instances simulated from the Mallows model with zero correlation disparity, \( \phi_\Delta = 0 \), symmetric correlation markets, and asymmetric correlation markets when simulated with \( \phi_\Delta \neq 0 \).

3 Stable Lattice under the Mallows model

In this section, we show empirically that in one-to-one markets with correlated preferences induced by Mallows models the size of the stable lattice depends on the correlation disparity. The empirical investigations give insights on how searching for a sex-equal matching can computationally vary with the size of a stable lattice in a given market.\(^4\)

---

\(^3\)Throughout the paper, we assume \( \hat{\pi}_m = \hat{\pi}_w \), because one can simply relabel the preferences on one side to create any arbitrary distance.

\(^4\)The source code is available at https://github.com/Restel/mallows-smp
| n  | Type  | Median $\mathcal{L}$ | CI($\mathcal{L}$) | max($\mathcal{L}$) |
|----|-------|----------------------|-------------------|-------------------|
| 1  | 150   | Mallows              | 16.00 (2;256)     | 1024              |
| 3  | 150   | Uniform              | 82.00 (32;231)    | 626               |
| 4  | 300   | Mallows              | 144.00 (8;7008)   | 147456            |
| 5  | 300   | Uniform              | 209.00 (91;567)   | 1262              |

Table 1: The size of the stable lattice ($\mathcal{L}$) under the Uniform, and the Mallows distributions with $\phi_m = \phi_w = 0.5$ models. $CI(\mathcal{L})$ denotes the 95% confidence interval.

3.1 Setup and Preference Sampling

We generate 1,000 instances of a stable matching problem for each $n \in [10, 150]$ and $\phi_m, \phi_w$ in $[0.1, 1.0]$. When generating a preference profile under probabilistic models, we draw a preference list for each agent i.i.d from the Mallows distribution using the sampler from PrefLib library [52].

In all sampled stable matching instances we built the stable lattice, and measured $S_M$, $S_W$ welfare scores. For each experimental setting, we estimate a target statistic (the median for numerical variables like size and mean for binary variables) over 1,000 instances along with the confidence intervals. Confidence intervals were obtained using bootstrap sampling with replacement with sample size 1,000 with 100 repeats. We report the results only for $n = 150$, as for the smaller $n$ the results are qualitatively the same. We occasionally considered larger instances with $n = 300$, when comparing the size of the lattice of the Uniform and the Mallows ($\phi_m = \phi_w = 0.5$) models.5

3.2 The Size of the Stable Lattice

The size of the stable lattice depends on the correlation disparity

Our empirical investigation illustrates an intriguing observation about the size of the stable lattice: Under the Mallows model the size of the stable lattice is affected by the correlation intensity in men’s and women’s preferences (measured by the dispersion parameters $\phi_m, \phi_w$) as well as the discrepancy between them, i.e. the correlation disparity $\phi_\Delta$. The size of the lattice is maximum when the preferences of men and women are sampled from a distribution with identical dispersion parameters, $\phi_\Delta = 0$, and decreases with the rise of the correlation disparity. This relationship is confirmed by median values, 25%-75% percentiles, 95% confidence intervals, maximum values and holds for all combinations of dispersion parameters (Figure 2). These findings are aligned with theoretical results indicating that when the preferences are sampled from the same distribution (parametrized by identical dispersion parameters, $\phi_m = \phi_w$), the size of the stable lattice grows exponentially in the number of agents [48], making the use of an exhaustive search algorithm impractical in these settings.

More importantly, they reveal that the primary predictor of the lattice size is the correlation disparity between the two sides, and the intensity of the correlations (dispersion parameters) have less significant impact on the lattice size.6

Mallows preferences induce large lattices

It is thought that introducing correlation in preferences reduces the stable lattice size compared to the random uncorrelated case [6, 47, 64]. In their empirical work, Roth and Peranson [64] conjectured that having

---

5 We use the same instance generation and setup in all remaining sections.

6 Similar results hold for rotation posets as we discuss in Appendix B.
a common objective ranking for the agents’ preferences is one of the main reason behind the observed small lattice sizes. Our empirical results suggest that while in general the median size of the lattice was indeed smaller in the Mallows compared to the Uniform distribution, some \( \phi \) parameters have much larger maximum values and confidence intervals.\(^7\) For example, in case of asymmetric correlation with \( \phi_m = 0.5, \phi_w = 0.3, n = 150 \), the maximum size of the lattice was larger than that of the Uniform model. Similarly, for symmetric case with \( \phi_m = 0.5, \phi_w = 0.5, \) and \( n = 300 \), 95% confidence interval of a stable lattice size spans from 8 to 7008 matchings compared to 91–567 interval under the Uniform distribution (see Table 1).

Our results suggest that symmetrically correlated markets have extraordinarily large lattices compared to the Uniform case (147456 vs. 1262), aligning well with the results of Levy [48] on asymptotic exponential lattices in the Mallow model. Note, that such large instances appeared for such small number of agents as 300, while real-world markets often have many more agents [57, 64].

In the past, correlation in preferences of agents was considered a sole factor contributing to small lattice sizes [6, 64]. Our experiments imply that in case of the correlation induced by the Mallows model, this effect depends not on the presence of correlation itself, but rather on its disparity between the two sides i.e. how strongly the preferences of one set are correlated compared to the preferences of the other set. Symmetric correlation markets seem to be the worst case scenario, and the lattice tends to grow smaller when the sets are having preferences correlated dissimilarly. This raises the question of whether real-world markets with observed small lattices suffer from such correlation disparity in preference profiles or it should be attributed to other factors.

4 Sex-Equal Stable Matchings

In this section, we discuss how the position of a sex-equal matching within a stable lattice is determined by the relationship between men and women welfare scores, which enables us to find a sex-equal matching using the DA algorithm.

The relationship between \( S_M \) and \( S_W \) of men- and women-optimal stable matchings dictates the location

\(^7\)See Table 4 in Appendix B for details.
of a sex-equal solution within the lattice. In particular, it determines whether or not it lies on the extreme points of the lattice. Based on this relationship, we identify two “easy” cases, in which DA guarantees to return a sex-equal stable matching, and a “hard” case, in which it does not.

**Lemma 1** (Kato [40]). A sex-equal stable matching is:

1) the men-optimal stable matching, $\mu^M$, if $S_M(\mu^M) \geq S_W(\mu^M)$

2) the women-optimal matching, $\mu^W$, if $S_M(\mu^W) \leq S_W(\mu^W)$

3) the stable matching $\mu_Z$ such that $c(\mu_Z) = \arg\min_{\mu \in S^>}(c(\mu))$, otherwise.

**Corollary 1.** A sex-equal stable matching can be computed in polynomial time when either $S_M(\mu^M) \geq S_W(\mu^M)$ or $S_M(\mu^W) \leq S_W(\mu^W)$.

**Proof.** Find a men-optimal matching $\mu^M$ by running a men-proposing DA algorithm. Compute $S_M(\mu^M)$ and $S_W(\mu^M)$ and if $S_M(\mu^M) \geq S_W(\mu^M)$, then by Lemma 1, Case 1 the men-optimal matching is a sex-equal stable matching. Analogously, find a women-optimal matching $\mu^W$ by running a women-proposing DA algorithm. Compute $S_M(\mu^W)$ and $S_W(\mu^W)$ and if $S_M(\mu^W) \leq S_W(\mu^W)$, then by Lemma 1, Case 2 the women-optimal matching is a sex-equal stable solution. The running time of DA and calculating $S_M$, $S_W$ scores is $O(n^2)$. Thus, a sex-equal stable matching can be found in polynomial time when $S_M(\mu^M) \geq S_W(\mu^M)$ or $S_M(\mu^W) \leq S_W(\mu^W)$.

In the first and the second cases of Lemma 1, the sign of the difference between the scores $S_M(\mu) - S_W(\mu)$ is preserved across the lattice, in other words $S_M(\mu)$ is either always greater or equal than $S_W(\mu)$ (Case 1), or smaller or equal than $S_W(\mu)$ (Case 2). The third case corresponds to the instances, in which the sign of $S_M(\mu) - S_W(\mu)$ changes across the lattice, and the location of the sex-equal matching within the lattice is not known a priori.

**4.1 Asymmetric Markets ($\phi_m \neq \phi_w$)**

When preferences of men and women have different levels of correlation intensity with the reference rankings, a sex-equal stable matching is often found on an extreme point of the stable lattice: men-optimal in case men have higher correlation and women-optimal otherwise. Empirically, we observe this behavior in all instances, including the largest stable lattices with up to 1024 matchings (Figure 3). We found that agents from the side with a smaller dispersion parameter $\phi$ of the Mallows model, are less satisfied with
their optimal matching than their partners with their pessimal one (and the market belongs to Case 1 or Case 2 scenario from Lemma 1). This fact might imply that, in asymmetric correlation markets, “hard” cases rarely occur and the DA algorithm can be considered a fair alternative to the exhaustive search, especially since such markets can induce large lattices as shown in Section 3.

Experimentally, in every stable matching instance belonging to an asymmetric correlation market with $\phi_m < \phi_w$, $S_M$-optimal was larger than $S_W$-pessimal, and therefore a men-optimal matching is sex-equal (Lemma 1, Case 1). Similarly, whenever $\phi_m > \phi_w$, $S_W$-optimal was larger than $S_M$-pessimal, rendering the women-optimal matching sex-equal (Lemma 1, Case 2). This is well illustrated by the distribution of $S_M$ and $S_W$ scores of men- and women-optimal matchings for various combinations of $\phi_m$, $\phi_w$. Figure 5 illustrates this relationship when men sample their preferences with $\phi_m = 0.9$, and women have their dispersion parameter in range $[0.1, 1.0]$ (the graphs for other values of $\phi_m$ are analogous). In the part of the graph, where women’s preferences have a lower dispersion parameter $\phi_w < \phi_m$, the distribution of $S_W$-scores of a women-optimal matching lies strictly above the distribution of its $S_M$-scores with no overlap.\(^8\)

![Figure 5: Illustrating $S_M, S_W$ scores under the Mallows model with $\phi_m = 0.9$ and varying $\phi_w$ based on 1000 samples.](image)

**Implications on the DA algorithm.**

Interestingly, increasing the variability of women’s preferences, results in 1) higher welfare for women in optimal and pessimal matchings and 2) a higher number of proposals from men. The total number of proposals in the DA algorithm corresponds to the optimal score of the proposing side [58]. In Figure 5 we can see a noticeable increase in the $S_M$-scores with the increase of $\phi_w$ as well as a decrease in $S_W$ scores (same pattern occurs for other values of $\phi_m$). Imagine that the proposing side samples its preferences from the model with a smaller dispersion parameter, and hence has less variable preferences and demonstrates a higher competition for partners (than the opposite side). Intuitively, during the course of DA those agents would propose to the same set of partners over and over again and their potential partners get to choose from

\(^8\)A more fine-grained analysis with $\phi_w \in \{0.9, 0.91, 0.92, \ldots, 0.99\}$ confirms the same finding (see Appendix C).
a diverse set of proposals, which improves the rank of their matches. This makes the total rank of proposers’ partners greater than that of the agents from the accepting side. This fact distinguishes the Mallows model from the Uniform model in which the proposing side gets a substantially smaller score than the accepting side [58].

4.2 Symmetric Markets \((\phi_m = \phi_w)\)

When the preferences of men and women have the same intensity of correlation, reflected in zero correlation disparity, the lattice size can quickly become very large. The size of the stable lattice is expected to be asymptotically exponential [48], which is confirmed in our experiments: even for small number of agents as \(n = 300\), there were significantly large (compared to the Uniform model) lattices with up to 147456 matchings (Table 1). Despite the large size of the stable lattice, the cost of a sex-equal solution is considerably close to the cost of DA outcomes (Figure 4). In this case, the DA algorithm can be considered cost-effective as the exhaustive search takes significantly more time for large lattices and gives only a moderate gain in the sex-equality cost.

We discussed in Section 4.1, that under the Mallows model, \(S_M\)-optimal and \(S_W\)-pessimal scores (similarly \(S_W\)-optimal, \(S_M\)-pessimal) are close to each other and their distributions overlap significantly. This can explain the small sex-equality cost of DA outcomes (see Figure 5). It could indicate, that under this scenario, the expected score of the proposing side under DA is similar to the expected score of the receivers. This finding differs sharply from the Uniform model wherein the expected pessimal score is substantially greater than the optimal one [7, 58]. As the difference between the welfare of men and women is sufficiently small, the importance of which group proposes (men or women) in DA becomes negligible. Hence, not only can the DA algorithm be adopted to ensure sex-equality, but it may also provide a natural framework for achieving procedural fairness as it has recently been studied [70, 71].

4.3 Implications on manipulation strategies

In our experiments, the Mallows model with symmetric and asymmetric correlations results in a considerably small difference between the pessimal and the optimal scores for one side (the welfare gap), for example, between \(S_W\) pessimal and \(S_W\)-optimal (Figure 5). This difference represents how much on average an agent can improve by shifting from her worst stable partner to her best one, and therefore has implications on manipulation incentives [6, 59]. Since the welfare gap is small for both men and women, in symmetric and asymmetric correlation settings, the accepting side has a limited scope/incentive for manipulations in practice. This observation suggests that one-to-one markets with Mallows preferences might be less susceptible to manipulation compared to the Uniform model, for which a larger welfare gap, and subsequently, higher incentive for manipulation has been established [58] and shown empirically [30, 69].

5 Comparing with Other Algorithms

We compare the DA algorithm with several state-of-the art algorithms (as implemented in [71]) such as multiple variants of the procedurally fair algorithms, and the best known heuristic algorithm when preferences are drawn from the Mallows model.\(^9\) We use an exhaustive search algorithm as our baseline to find sex-equal stable matchings [70]. In particular, we test several procedures allowing both sides to issue and

\(^9\)The comparison of algorithms for the Uniform model can be found in Appendix E.
receive proposals such as Late Discontent Dispension (LDS), Early Discontent Dispension, and Powerbalance [71]. We also measure the performance of an enhanced bidirectional local search iBILS, which is one of the best-performing existing heuristic for finding sex-equal stable matchings.

5.1 Experimental Setup

We evaluate the sex-equality cost and the run-time of these algorithms for the number of agents varying from 20 to 1000. For the symmetric correlation case, we select \( \phi_m = \phi_w = 0.5 \) as these parameters result in the largest lattice size among all combinations of \( \phi_m \) and \( \phi_w \) (see Table 4). We sample 1000 instances of the stable matching for each experimental setting.

The exhaustive search algorithm uses the break-marriage operation of McVitie and Wilson [54]: breaking a single man’s assignment initiates a chain of proposals and rejections. It terminates when either all women reject the last man, or his proposal is accepted by a single woman. The latter case results in a new stable matching. All the elements of the stable lattice can be found by applying this operation recursively from the men-optimal solution [54].\(^{10}\)

The enhanced bidirectional local search, iBILS, uses rotations to generate the successors of a stable matching to search over the stable lattice starting from its extreme points [70, 73].

5.2 Fairness and Running Time

Symmetric correlation. For symmetric correlation markets we used a smaller number of agents (up to 150), to avoid instances with extremely large lattices, for which exhaustive search would not terminate in a reasonable time (Section 4.2). Under symmetric correlations, stable lattice can grow exponentially [48]. In our experiments, all algorithms performed similarly on average with respect to the sex-equality cost (Figure 6). Among them, the iBILS algorithm performed better, and in most instances, found the sex-equal stable matching that was computed by the exhaustive search algorithm. These results and the good performance of other algorithms are justified by the fact that under the Mallows model, \( S_M \) and \( S_W \) scores of all stable matchings are close to one another (i.e. the welfare gap between the two sets is small). Therefore, many stable matchings have costs close to that of the the sex-equal solution.

\(^{10}\)An example of the break-marriage operation is provided in Appendix A.
While the DA and procedurally fair algorithms give higher sex-equality costs (on average twice as much as the exhaustive search and the iBILS algorithms), they have much smaller running times. For larger instances, this discrepancy between the running times will become even more pronounced. The exhaustive search has exponential complexity in the worst case. The running time of the iBILS algorithm depends on the stable lattice size and its configuration: it is $O(dkn^2)$, where $d$ is the maximum search depth and $k$ is the maximum width of the lattice [70, 74]. The DA algorithm and procedurally fair algorithms run in polynomial-time.

**Asymmetric correlation.** In asymmetric correlation markets, we focused on comparing the performance of iBILS and DA with additional pre-processing step (aka DA*). Given the findings in Section 4.1, we modify the DA algorithm by deciding which side should propose according to the dispersion parameters: the side with smaller dispersion parameter proposes, i.e. when $\phi_m < \phi_w$ men propose, otherwise women propose. In real-world markets, dispersion parameters can be inferred from preference profiles in polynomial time when the reference ranking is known [33].

Figure 7 presents the sex-equality cost of DA* compared to the iBILS algorithm for large markets with 300 and 1000 agents. In almost all instances, DA* finds the same stable matching as iBILS. More importantly, as shown in Figure 8, the run-time of the DA* remains in $O(n^2)$ while iBILS requires more steps to estimate a sex-equal solution. Given the performance of iBILS in most cases, our results suggest that DA* is a reasonable cost-effective fair algorithm for matching markets with asymmetric correlations.

![Figure 8](image.png)

**Figure 8:** Mallows model (asymmetric): the running time of iBILS and a pre-processed Deferred Acceptance (DA*) for an asymmetric correlation market with $\phi_m = 0.5, \phi_w = 0.7$ (other combinations of parameters give analogous results).

6 Concluding Remarks

For decades, market design has played a crucial role in setting the foundations of decision-making through mechanism design. While the theoretical guarantees sometimes fall short due to incompatibilities between axiomatic requirements or challenges imposed by computational complexity, empirical evaluations can still shed light on practical implications and guide the design of new mechanisms. This in turn could help produce new problems and directions for traditional theory.

In this spirit, our extensive empirical investigations revealed an intriguing relation between the disparity of preferences of both sides and the fairness of stable matchings. We showed that surprisingly, the primary
factor affecting the sex-equality cost of the DA algorithm is the difference between dispersion parameters of both sides. These observations suggest that the DA algorithm could still be a prime candidate both in symmetric and asymmetric markets, further justifying its wide use in practice.

From the theoretical perspective, investigating the bounds of fairness in stable matchings under correlated preferences is certainly an interesting future direction. From the practical perspective, a noteworthy, and perhaps more important, direction is to investigate the structure of preferences in various matching settings (e.g. refugee matching, residency matching, and school choice) and develop domain-specific models that can correctly capture the correlations between the preferences with high accuracy. We believe that these models, alongside with theoretical studies, should shape governmental and societal policies—prescribed by social planners—in the adoption of suitable mechanisms in each specific domain.

Acknowledgments

Hadi Hosseini acknowledges support from NSF IIS grants #2052488 and #2107173. We are grateful to the anonymous referees for their valuable comments and suggestions.

References

[1] Atila Abdulkadiroğlu and Tayfun Sonsmez. School choice: A mechanism design approach. *American Economic Review*, 93(3):729–747, 2003.

[2] Atila Abdulkadiroğlu, Parag A Pathak, and Alvin E Roth. The New York City High School Match. *American Economic Review*, 95(2):364–367, 2005.

[3] Atila Abdulkadiroğlu, Parag A Pathak, and Alvin E Roth. The new york city high school match. *American Economic Review*, 95(2):364–367, 2005.

[4] Atila Abdulkadiroğlu, Parag A Pathak, Alvin E Roth, and Tayfun Sonsmez. The Boston Public School Match. *American Economic Review*, 95(2):368–371, 2005.

[5] Brian Aldershof, Olivia M Carducci, and David C Lorenc. Refined inequalities for stable marriage. *Constraints*, 4(3):281–292, 1999.

[6] Itai Ashlagi, Yash Kanoria, and Jacob D Leshno. Unbalanced random matching markets: The stark effect of competition. *Journal of Political Economy*, 125(1):69–98, 2017.

[7] Itai Ashlagi, Mark Braverman, Clayton Thomas, and Geng Zhao. Tiered random matching markets: Rank is proportional to popularity. *arXiv preprint arXiv:2009.05124*, 2020.

[8] Pranjal Awasthi, Avrim Blum, Or Sheffet, and Aravindan Vijayaraghavan. Learning mixtures of ranking models. In *Advances in Neural Information Processing Systems*, pages 2609–2617, 2014.

[9] Siddhartha Banerjee and Ramesh Johari. Ride sharing. In *Sharing economy*, pages 73–97. Springer, 2019.

[10] Hedeyeh Beyhaghi, Daniela Saban, and Eva Tardos. Effect of selfish choices in deferred acceptance with short lists. *arXiv preprint arXiv:1701.00849*, 2017.
[11] Nayantara Bhatnagar, Sam Greenberg, and Dana Randall. Sampling stable marriages: why spouse-swapping won’t work. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1223–1232. Society for Industrial and Applied Mathematics, 2008.

[12] Duncan Black et al. The theory of committees and elections. 1958.

[13] James W Boudreau and Vicki Knoblauch. Marriage matching and intercorrelation of preferences. *Journal of Public Economic Theory*, 12(3):587–602, 2010.

[14] Frank A Campo and Marcel Erné. Exponential functions of finite posets and the number of extensions with a fixed set of minimal points. *arXiv preprint arXiv:1802.01419*, 2018.

[15] Ioannis Caragiannis, Ariel D Procaccia, and Nisarg Shah. When do noisy votes reveal the truth? In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, pages 143–160, 2013.

[16] Onur B Celik and Vicki Knoblauch. Marriage matching with correlated preferences. 2007.

[17] Christine T Cheng. Understanding the generalized median stable matchings. *Algorithmica*, 58(1):34–51, 2010.

[18] Christine T Cheng and Will Rosenbaum. Stable matchings with restricted preferences: Structure and complexity. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 319–339, 2021.

[19] David Delacretaz, Scott Kominers, and Alexander Teytelboym. Matching mechanisms for refugee resettlement. Technical report, Human Capital and Economic Opportunity Working Group, 2019.

[20] Gabrielle Demange, David Gale, and Marilda Sotomayor. A further note on the stable matching problem. *Discrete Applied Mathematics*, 16(3):217–222, 1987.

[21] Tomás Feder. A new fixed point approach for stable networks and stable marriages. *Journal of Computer and System Sciences*, 45(2):233–284, 1992.

[22] Rupert Freeman, Evi Micha, and Nisarg Shah. Two-sided matching meets fair division.

[23] David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

[24] David Gale and Marilda Sotomayor. Some Remarks on the Stable Matching Problem. *Discrete Applied Mathematics*, 11(3):223–232, 1985.

[25] Sushmita Gupta, Saket Saurabh, and Meirav Zehavi. On treewidth and stable marriage. *arXiv preprint arXiv:1707.05404*, 2017.

[26] Dan Gusfield. Three fast algorithms for four problems in stable marriage. *SIAM Journal on Computing*, 16(1):111–128, 1987.

[27] Dan Gusfield and Robert W Irving. *The stable marriage problem: structure and algorithms*. MIT press, 1989.

[28] Jonathan V Hall, John J Horton, and Daniel T Knoepfle. Labor market equilibration: Evidence from Uber. 2017. Working paper. Revision requested from American Economic Review.
[29] Avinatan Hassidim, Assaf Romm, and Ran I Shorrer. Need vs. merit: The large core of college admissions markets. *Merit: The Large Core of College Admissions Markets (June 28, 2018)*, 2018.

[30] Hadi Hosseini, Fatima Umar, and Rohit Vaish. Accomplice Manipulation of the Deferred Acceptance Algorithm. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pages 231–237, 2021.

[31] Chien-Chung Huang. Classified stable matching. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 1235–1253. SIAM, 2010.

[32] Nicole Immorlica and Mohammad Mahdian. Incentives in large random two-sided markets. *ACM Transactions on Economics and Computation (TEAC)*, 3(3):1–25, 2015.

[33] Ekhine Irurozki. *Sampling and learning distance-based probability models for permutation spaces*. PhD thesis, 2014.

[34] Ekhine Irurozki, Borja Calvo, Jose A Lozano, et al. Mallows and generalized mallows model for matchings. *Bernoulli*, 25(2):1160–1188, 2019.

[35] Robert W Irving and Paul Leather. The complexity of counting stable marriages. *SIAM Journal on Computing*, 15(3):655–667, 1986.

[36] Robert W Irving, Paul Leather, and Dan Gusfield. An efficient algorithm for the “optimal” stable marriage. *Journal of the ACM (JACM)*, 34(3):532–543, 1987.

[37] Robert W Irving, David F Manlove, and Sandy Scott. The stable marriage problem with master preference lists. *Discrete Applied Mathematics*, 156(15):2959–2977, 2008.

[38] Will Jones and Alexander Teytelboym. Matching systems for refugees. *Journal on Migration and Human Security*, 5(3):667–681, 2017. doi: 10.1177/233150241700500306. URL https://doi.org/10.1177/233150241700500306.

[39] Yuichiro Kamada and Fuhito Kojima. Efficient matching under distributional constraints: Theory and applications. *American Economic Review*, 105(1):67–99, 2015.

[40] Akiko Kato. Complexity of the sex-equal stable marriage problem. *Japan Journal of Industrial and Applied Mathematics*, 10(1):1–19, 1993.

[41] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[42] MG Kendall et al. Rank correlation methods. *Rank correlation methods.*, 1948.

[43] Bettina Klaus and Flip Klijn. Procedurally fair and stable matching. *Economic Theory*, 27(2):431–447, 2006.

[44] Vicki Knoblauch. Marriage matching and gender satisfaction. *Social Choice and Welfare*, 32(1):15–27, 2009.

[45] Donald E Knuth, Rajeev Motwani, and Boris Pittel. Stable husbands. *Random Structures & Algorithms*, 1(1):1–14, 1990.
[46] Fuhito Kojima and Parag A Pathak. Incentives and stability in large two-sided matching markets. *American Economic Review*, 99(3):608–27, 2009.

[47] SangMok Lee. Incentive compatibility of large centralized matching markets. *The Review of Economic Studies*, 84(1):444–463, 2016.

[48] Avi William Levy. *Novel uses of the Mallows model in coloring and matching*. PhD thesis, Dept. of Mathematics, University of Washington, Seattle, 2017.

[49] Tyler Lu and Craig Boutilier. Effective sampling and learning for mallows models with pairwise-preference data. *The Journal of Machine Learning Research*, 15(1):3783–3829, 2014.

[50] Colin L Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.

[51] John I Marden. Analyzing and modeling rank data. 1995.

[52] Nicholas Mattei and Toby Walsh. Preflib: A library of preference data HTTP://PREFLIB.ORG. In *Proceedings of the 3rd International Conference on Algorithmic Decision Theory (ADT 2013)*, Lecture Notes in Artificial Intelligence. Springer, 2013.

[53] Eric McDermid and Robert W Irving. Sex-equal stable matchings: Complexity and exact algorithms. *Algorithmica*, 68(3):545–570, 2014.

[54] David G McVitie and Leslie B Wilson. The stable marriage problem. *Communications of the ACM*, 14(7):486–490, 1971.

[55] Morikazu Nakamura, Kenji Onaga, Seiki Kyan, and Manuel Silva. Genetic algorithm for sex-fair stable marriage problem. In *Proceedings of ISCAS’95-International Symposium on Circuits and Systems*, volume 1, pages 509–512. IEEE, 1995.

[56] Thanh Nguyen and Rakesh Vohra. Stable matching with proportionality constraints. *Operations Research*, 67(6):1503–1519, 2019.

[57] Parag Pathak and Tayfun Sönmez. c. 2007.

[58] Boris Pittel. The average number of stable matchings. *SIAM Journal on Discrete Mathematics*, 2(4):530–549, 1989.

[59] Ross Rheingans-Yoo and Jane Street. Large random matching markets with localized preference structures can exhibit large cores. Technical report, Mimeo, 2020.

[60] Antonio Romero-Medina. Equitable selection in bilateral matching markets. *Theory and Decision*, 58(3):305–324, 2005.

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

[62] Alvin 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.

[63] Alvin E Roth. The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, 70(4):1341–1378, 2002.
[64] Alvin E Roth and Elliott Peranson. The redesign of the matching market for american physicians: Some engineering aspects of economic design. *American economic review*, 89(4):748–780, 1999.

[65] Alvin E Roth and John H Vande Vate. Random paths to stability in two-sided matching. *Econometrica: Journal of the Econometric Society*, pages 1475–1480, 1990.

[66] Hannu Salonen and Mikko AA Salonen. Mutually best matches. *Mathematical Social Sciences*, 91:42–50, 2018.

[67] Weiran Shen, Pingzhong Tang, and Yuan Deng. Coalition manipulation of gale-shapley algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[68] Mark Steyvers, Brent Miller, Pernille Hemmer, and Michael D Lee. The wisdom of crowds in the recollection of order information. In *Advances in Neural Information Processing Systems*, pages 1785–1793, 2009.

[69] Chung-Piaw Teo, Jay Sethuraman, and Wee-Peng Tan. Gale-shapley stable marriage problem revisited: Strategic issues and applications. *Management Science*, 47(9):1252–1267, 2001.

[70] Nikolaos Tziavelis, Ioannis Giannakopoulos, Katerina Doka, Nectarios Koziris, and Panagiotis Karras. Equitable stable matchings in quadratic time. In *Advances in Neural Information Processing Systems*, pages 455–465, 2019.

[71] Nikolaos Tziavelis, Ioannis Giannakopoulos, Rune Quist Johansen, Katerina Doka, Nectarios Koziris, and Panagiotis Karras. Fair procedures for fair stable marriage outcomes. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[72] Rohit Vaish and Dinesh Garg. Manipulating gale-shapley algorithm: Preserving stability and remaining inconspicuous. In *IJCAI*, pages 437–443, 2017.

[73] Hoang Huu Viet, SeungGwan Lee, Taechoong Chung, et al. A bidirectional local search for the stable marriage problem. In *2016 International Conference on Advanced Computing and Applications (ACOMP)*, pages 18–24. IEEE, 2016.

[74] Hoang Huu Viet, Le Hong Trang, Le Pham Tuyen, and Taechoong Chung. A shortlist-based bidirectional local search for the stable marriage problem. *Journal of Experimental & Theoretical Artificial Intelligence*, 32(1):147–163, 2020.

[75] Valeria Vitelli, Øystein Sørensen, Marta Crispino, Arnoldo Frigessi, and Elja Arjas. Probabilistic preference learning with the mallows rank model. *The Journal of Machine Learning Research*, 18(1):5796–5844, 2017.
Supplementary Material

A Additional Preliminaries

In this section, we give additional preliminaries, definitions, and examples to provide a complete view to the readers.

Dominance

Stable matching \( \mu' \) dominates stable matching \( \mu \) if it assigns all men a partner at least as good as in \( \mu \) and a better partner to at least one man [27]. Formally, \( \mu' \) dominates \( \mu \) if \( \forall m \in M \ r(\mu'(m), m) \leq r(\mu(m), m) \) and \( \exists m^* \in M \) such that \( r(\mu'(m^*), m^*) < r(\mu(m^*), m^*) \).

Stable lattice

Given a preference profile \( \succ \), the set of all corresponding stable matchings, \( S_\succ \), forms a distributive lattice under a dominance relation (referred to as a stable lattice hereinafter). Any pair of stable matchings in a stable lattice has a unique least upper bound (also called a join) and a unique greatest lower bound (a meet) within the lattice. In a join of two stable matchings, \( \mu_1, \mu_2 \), every man from \( M \) is given its best partner among \( \mu_1, \mu_2 \). Similarly, in a meet all men are given their worse partners from \( \mu_1, \mu_2 \).

In a stable lattice supremum, the men-optimal matching \( \mu^M \), all men are matched to their favorite partners among all potential stable partners. It dominates every other matching in \( S_\succ \), formally \( \forall m \in M \ r(\mu^M(m), m) \leq r(\mu'(m), m) \) for all \( \mu' \in S_\succ \). Similarly, in a stable lattice infimum, the men-pessimal matching \( \mu^W \), all men are given their worse stable partners. Men-pessimal matching is dominated by every other stable matching. Interestingly, a stable lattice infimum is simultaneously women-optimal matching: each woman is matched to her best stable partner in \( S_\succ \), i.e., \( \forall w \in W \ r(\mu^W(w), w) \leq r(\mu'(w), w) \) for all \( \mu' \in S_\succ \). Figure 1 illustrates a lattice consisting of six stable matchings.

Rotation poset.

For any stable lattice \( S_\succ \) there exists an associated partially ordered set (a rotation poset, \( R(S_\succ) \)), whose downsets are in one-to-one correspondence with the matchings of the lattice [27]. The elements of the rotation poset—aka rotations—are specific sequences of pairs matched in some stable matching. For each man \( m \) in a rotation the following holds: 1) the woman in his pair is his current partner \( \mu(m) \); 2) the woman to the right of \( m \) (or the first woman if \( m \) is the last man), \( w_{\text{next}}(m) \), would be the first to accept his proposal if he is disallowed to propose to \( \mu(m) \) within the course of \( D\mu \). Formally, a rotation \( \rho = (m_1, w_1), (m_2, w_2), ..., (m_k, w_k) \) is a sequence of pairs in a matching such that \( m_i \) is matched to \( w_i \), and \( w_{(i+1) \bmod k} \) is the first woman in \( m_i \)’s preference list who prefers him to her current partner. For specific \( \mu \) and \( \rho \), if for every man \( m_i \in \rho \) the above conditions holds, the rotation \( \rho \) is said to be exposed in \( \mu \). One rotation can be exposed in multiple stable matchings of the same problem instance; a stable matching can have multiple exposed rotations [27].

Given \( \rho \) exposed in \( \mu \), if every man \( m_i \) from the rotation is matched to the woman to the right of him, \( w_{i+1} \) (where \( (i + 1) \) is taken modulo size of \( \rho \)), and every man not in the rotation is assigned his current partner, then the resulting matching is also stable [35]. Rotations can be thought of as the transition operations between the adjacent elements of the lattice. The rotation poset is a useful structure for studying the stable matching problem [11, 17, 26] as it contains \( O(n^2) \) elements (in contrast to the exponential size
her partner along with his current partner. Every woman whose current partner is the last entry in her preference list, to her partner, puts all men, that she does not prefer over her current partner, i.e. \( w \) next \( \notin W \), from her list all men, that she does not prefer over her current partner, i.e. \( w \). Removed men delete woman \( w \) from their preferences as well. In the resulting male-oriented shortlist, the following holds for every \( m \in \mu \): 1) \( m \)'s current partner is the first entry in his list \( \succ_m \); 2) the first woman who prefers \( m \) to her partner, \( w_{\text{next}}(m) \), is the second entry in \( \succ_m \). Also, for every woman \( w \in W \) her current partner is the last entry in her preference list \( \succ_w \) [36].

To identify an exposed rotation, we will take the first man \( m_1 \) and put it in front of a candidate rotation along with his current partner \( \rho_1 = (m_1, \mu(m_1)) \). The next pair in the sequence must contain \( w_{\text{next}}(m_1) \) and her partner \( \mu(w_{\text{next}}(m_1)) \), i.e. \( \rho_1 = (m_1, \mu(m_1)), (\mu(w_{\text{next}}(m_1)), w_{\text{next}}(m_1)) \). We will repeat building the candidate rotation in this fashion. The rotation becomes valid as soon as one of the men appears twice in the same candidate rotation. After this happens, we will discard all pairs outside of the cycle: the remaining ordered sequence of pairs constitutes a rotation exposed in \( \mu \). In order to identify all rotations exposed in the same matching, the same procedure must be repeated for all men that did not appear in a valid rotation and were not discarded in the course of building one.

We will identify all rotations exposed in \( \mu^M \) to obtain several stable matchings adjacent to it. We start

\[
\begin{align*}
&\left( w_2, w_3, w_1, w_4, w_5 \right), \text{ Men-optimal} \\
&\left( w_4, w_2, w_5, w_3, w_1 \right), \text{ Women-optimal (Sex-equal)}
\end{align*}
\]

Figure 9: The stable lattice for the profile described in Example 1.1 with six stable solutions. Matchings are denoted as lists of women matched to men according to their indices, rotations are indicated by \( \rho_\mu \) on arcs.

\[
\begin{align*}
&\mu^M = (w_2, w_3, w_1, w_4, w_5), \\
&w_1 : m_4 \quad m_2 \quad m_1 \quad m_5 \quad m_3^\dagger \\
&m_2 : w_3^\dagger \quad w_2 \quad w_4 \quad w_1 \quad w_5 \\
&m_3 : w_1^\dagger \quad w_5 \\
&m_4 : w_2^\dagger \quad w_3 \quad w_1 \quad w_5 \\
&m_5 : w_3 \quad w_1 \\
&\rho_1 = (m_1, \mu(m_1)), (\mu(w_{\text{next}}(m_1)), w_{\text{next}}(m_1))
\end{align*}
\]

Table 2: The shortlist of men-optimal matching of Example 1.1 instance

of the stable lattice in the worst case) and yet encodes all stable matchings. We refer the reader to [27] for a more detailed discussion on the subject of rotations.

**Example from Figure 1: Stable Lattice and Rotations**

In this section, we illustrate the concept of stable lattice, rotation poset and break-marriage operation using the preference profile given in Example 1.1.

We run a men-proposing \( DA \) to obtain the men-optimal matching, \( \mu^M = (w_2, w_3, w_1, w_4, w_5) \), where women are matched to men according to their position in the matching, e.g. \( w_2 \) is matched to \( m_1, w_3 \) to \( m_2 \) and so on.

To identify a rotation exposed in \( \mu^M \) we will use the concept of shortlists [36]. Given a stable matching \( \mu \), let each woman \( w \in W \) remove from her list all men, that she does not prefer over her current partner, i.e. every \( m' \) s.t. \( \mu(m) \succ_w m' \). Removed men delete woman \( w \) from their preferences as well. In the resulting male-oriented shortlist, the following holds for every \( m \in \mu \): 1) \( m \)'s current partner is the first entry in his list \( \succ_m \); 2) the first woman who prefers \( m \) to her partner, \( w_{\text{next}}(m) \), is the second entry in \( \succ_m \). Also, for every woman \( w \in W \) her current partner is the last entry in her preference list \( \succ_w \) [36].

To identify an exposed rotation, we will take the first man \( m_1 \) and put it in front of a candidate rotation along with his current partner \( \rho_1 = (m_1, \mu(m_1)) \). The next pair in the sequence must contain \( w_{\text{next}}(m_1) \) and her partner \( \mu(w_{\text{next}}(m_1)) \), i.e. \( \rho_1 = (m_1, \mu(m_1)), (\mu(w_{\text{next}}(m_1)), w_{\text{next}}(m_1)) \). We will repeat building the candidate rotation in this fashion. The rotation becomes valid as soon as one of the men appears twice in the same candidate rotation. After this happens, we will discard all pairs outside of the cycle: the remaining ordered sequence of pairs constitutes a rotation exposed in \( \mu \). In order to identify all rotations exposed in the same matching, the same procedure must be repeated for all men that did not appear in a valid rotation and were not discarded in the course of building one.

We will identify all rotations exposed in \( \mu^M \) to obtain several stable matchings adjacent to it. We start
a candidate rotation with putting $m_1$ and his partner $w_2$ to the beginning: $\rho_1 = (m_1, w_2)$. The first woman who prefers $m_1$ over her current partner is $w_4$, and we append her and her partner as the next pair in the sequence: $\rho_1 = (m_1, w_2)(m_4, w_4)$. The next pair in the sequence will be $(m_1, w_2)$ as $w_{next}(m_4) = w_2$. As it is the second appearance of $m_1$ in the sequence we found a rotation exposed in $\mu^M$. We continue in this fashion with the next man $m_3$ as a founder of a new candidate rotation, that leads to $\rho_3 = (m_3, w_1), (m_5, w_5)$. Using $m_2$ as the first man of a rotation leads to already found $\rho_1$.

When rotation $\rho_1$ is applied to $\mu^M$, $m_1$ is matched to $w_4$, $m_4$ to $w_2$, giving stable matching $\mu_1 = (w_4, w_3, w_1, w_2, w_5)$. Similarly, the application of $\rho_3$ gives $\mu_2 = (w_2, w_3, w_5, w_4, w_1)$. By finding and applying rotations exposed in $\mu_1$, $\mu_2$ one can recreate the lattice from Example 1.1 with 6 elements.

One can also use a break-marriage algorithm to find all the elements of the lattice: it is a recursive procedure that starts with breaking assignments of the fist man in the men-optimal matching, and letting him propose to the next woman who has not rejected him yet. This initial change in the matching initiates a chain of proposals and rejections between the agents and terminates when either all women reject some man (the last one to propose), or his proposal is accepted by a single woman. The latter case results in a new stable matching. By applying this operation recursively to all new stable matchings one can find all the elements of the stable lattice [54].

Let apply break-marriage to the fist man in the men-optimal matching $\mu^M$: $m_1$ is assigned free and propose to his next favorite woman in the list $w_4$. She prefers his proposal to her current partner and is matched to $m_1$ tentatively. Man $m_4$ proposes to $w_2$, who accepts the proposal since she is currently single. The new matching is $\mu_1$ and the current breakmarriage round will now assign $m_2$ in $\mu^M$ free and tries to identify a new stable matching. If breakmarriage is applied to all men in the matching, it will result in repetitive stable matchings, so a more efficient modification of the algorithm exists [54].
### B Summary Statistics for Stable Lattice and Rotation Poset

#### Rotation Poset

The rotation poset size follows the same pattern as the lattice size: it has the maximum size in case of zero correlation disparity and decreases with the increase in the disparity. The rotation poset is much smaller under the Mallows model (all cases) compared to the Uniform. For \( n = 150 \), the median rotation poset size in the Uniform model (33) is substantially larger than in the Mallows model (maximum median equals 4 rotations) (Table 3). For \( n = 300 \), the difference is even more pronounced: 60 rotations vs 8. For a general case, if rotation poset \( \mathcal{R} \) has \( r \) rotations and height \( h \), the maximum number of downsets in a poset, \( d(\mathcal{P}) \), (corresponding to maximum stable lattice size) can be counted as \( d(\mathcal{P}) = 2^{r-h}(h+1) \) [14]. We see an interesting configuration of the rotation posets associated with large lattices: a decreased height (the size of the maximal chain), and an increased width (the size of the largest anti-chain) [14]. For example, for \( n = 150 \) the maximum stable lattice under the Mallows model has 2304 matchings, only 12 corresponding rotations and rotation poset height equals 2 (Table 4). Under the Uniform model the maximum stable lattice was smaller but had many more rotations: 626 matchings and 53 rotations (the rotation poset height was equal to 4).

### C Fine-grained asymmetric correlation case

In the part of the graph, where women’s preferences have a lower dispersion parameter \( \phi_w < \phi_m \), the distribution of \( S_W \)-scores of a women-optimal matching lies above the distribution of its \( S_M \)-scores: the distributions have no overlap for \( \phi_w \leq 0.96 \) (Figure 11).

### D Egalitarian cost under the Mallows model

To assess the total welfare of \( DA \) outcome in the presence of correlation, we looked into its egalitarian cost. We found that introducing correlation in preferences of any side translates into a high total dissatisfaction of all agents by the \( DA \) outcome.

Given a preference profile \( > \), the egalitarian cost of a stable matching is the total ranking over all men and women [36]. Formally,

\[
c'(\mu) = S_M(\mu) + S_W(\mu)
\]

The smaller the egalitarian cost, the higher the welfare of all involved agents.

According to our experiments, \( DA \) algorithm is a reasonable alternative to the exhaustive search algorithm for finding a sex-equal stable matching, when the preferences come from the Mallows model. Its output seems to be either a sex-equal solution (in asymmetric correlation markets) or a matching close to
| $\phi_w$ | $Q2(L)$ | $Q1(L)$ | $Q3(L)$ | max($L$) | max($R$) | $Q2(R)$ | $Q1(R)$ | $Q3(R)$ |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| $\phi_m = 0.1$ |
| 1 | 0.1 | 2.00 | 1.00 | 4.00 | 128 | 7 | 1.00 | 0.00 | 2.00 |
| 2 | 0.3 | 4.00 | 2.00 | 8.00 | 128 | 7 | 2.00 | 1.00 | 3.00 |
| 3 | 0.5 | 2.00 | 2.00 | 4.00 | 32 | 5 | 1.00 | 1.00 | 2.00 |
| 4 | 0.7 | 2.00 | 1.00 | 2.00 | 32 | 5 | 1.00 | 0.00 | 1.00 |
| 5 | 0.9 | 1.00 | 1.00 | 1.00 | 8 | 3 | 0.00 | 0.00 | 0.00 |
| 6 | 1 | 1.00 | 1.00 | 1.00 | 4 | 2 | 0.00 | 0.00 | 0.00 |
| $\phi_m = 0.3$ |
| 7 | 0.1 | 4.00 | 2.00 | 6.50 | 128 | 7 | 2.00 | 1.00 | 3.00 |
| 8 | 0.3 | 8.00 | 4.00 | 16.00 | 2048 | 11 | 3.00 | 2.00 | 4.00 |
| 9 | 0.5 | 8.00 | 4.00 | 16.00 | 768 | 10 | 3.00 | 2.00 | 4.00 |
| 10 | 0.7 | 4.00 | 2.00 | 8.00 | 256 | 8 | 2.00 | 1.00 | 3.00 |
| 11 | 0.9 | 1.00 | 1.00 | 2.00 | 16 | 4 | 0.00 | 0.00 | 1.00 |
| 12 | 1 | 1.00 | 1.00 | 2.00 | 8 | 3 | 0.00 | 0.00 | 1.00 |
| $\phi_m = 0.5$ |
| 13 | 0.1 | 2.00 | 1.00 | 4.00 | 64 | 6 | 1.00 | 0.00 | 2.00 |
| 14 | 0.3 | 8.00 | 4.00 | 16.00 | 1024 | 10 | 3.00 | 2.00 | 4.00 |
| 15 | 0.5 | 16.00 | 6.00 | 32.00 | 1024 | 11 | 4.00 | 3.00 | 5.00 |
| 16 | 0.7 | 8.00 | 4.00 | 16.00 | 512 | 9 | 3.00 | 2.00 | 4.00 |
| 17 | 0.9 | 2.00 | 1.00 | 4.00 | 96 | 7 | 1.00 | 0.00 | 2.00 |
| 18 | 1 | 1.00 | 1.00 | 2.00 | 12 | 5 | 0.00 | 0.00 | 1.00 |
| $\phi_m = 0.7$ |
| 19 | 0.1 | 2.00 | 1.00 | 2.00 | 16 | 4 | 1.00 | 0.00 | 1.00 |
| 20 | 0.3 | 4.00 | 2.00 | 8.00 | 192 | 8 | 2.00 | 1.00 | 3.00 |
| 21 | 0.5 | 8.00 | 4.00 | 16.00 | 384 | 10 | 3.00 | 2.00 | 4.00 |
| 22 | 0.7 | 12.00 | 4.00 | 32.00 | 2304 | 12 | 4.00 | 2.00 | 5.00 |
| 23 | 0.9 | 4.00 | 2.00 | 8.00 | 192 | 9 | 2.00 | 1.00 | 3.00 |
| 24 | 1 | 2.00 | 1.00 | 4.00 | 14 | 6 | 1.00 | 0.00 | 2.00 |
| $\phi_m = 0.9$ |
| 25 | 0.1 | 1.00 | 1.00 | 1.00 | 8 | 3 | 0.00 | 0.00 | 0.00 |
| 26 | 0.3 | 1.00 | 1.00 | 2.00 | 16 | 5 | 0.00 | 0.00 | 1.00 |
| 27 | 0.5 | 2.00 | 1.00 | 4.00 | 96 | 7 | 1.00 | 0.00 | 2.00 |
| 28 | 0.7 | 4.00 | 2.00 | 8.00 | 96 | 8 | 2.00 | 1.00 | 3.00 |
| 29 | 0.9 | 8.00 | 4.00 | 16.00 | 280 | 14 | 4.00 | 2.00 | 5.00 |
| 30 | 1 | 8.00 | 4.00 | 12.00 | 112 | 15 | 4.00 | 3.00 | 6.00 |
| $\phi_m = 1.0$ (Uniform) |
| 31 | 0.1 | 1.00 | 1.00 | 1.00 | 4 | 2 | 0.00 | 0.00 | 0.00 |
| 32 | 0.3 | 1.00 | 1.00 | 2.00 | 8 | 3 | 0.00 | 0.00 | 1.00 |
| 33 | 0.5 | 1.00 | 1.00 | 2.00 | 9 | 5 | 0.00 | 0.00 | 1.00 |
| 34 | 0.7 | 2.00 | 1.00 | 4.00 | 30 | 7 | 1.00 | 0.00 | 2.00 |
| 35 | 0.9 | 7.00 | 4.00 | 12.00 | 116 | 15 | 5.00 | 3.00 | 7.00 |
| 36 | 1 | 82.00 | 61.00 | 112.00 | 792 | 53 | 33.00 | 29.00 | 37.00 |

Table 4: Summary statistics for the size of the lattice ($L$) and the rotation poset ($R$) in Mallows model with different dispersion parameters for men ($\phi_m$) and women ($\phi_w$). Q1, Q2, Q3 indicate 25th, 50th (median), 75th percentiles.
sex-equal one in terms of the cost (in symmetric correlation markets). Although DA achieves such a high level of sex-equality fairness in the Mallows, the overall satisfaction of agents by its output remains low in experiments (Figure 12). The egalitarian cost is high for all \( \phi \) combinations under the Mallows distribution: the smallest median score is 16636 for \( \phi_m = \phi_w = 0.9 \), which for \( n = 150 \) corresponds to having a partner with 55 rank position on average. In contrast, the DA applied to the uniform preferences have a relatively high total satisfaction, while having a great disparity between the welfare of men and women according to our experiments (Figure 4).

Figure 12: The egalitarian cost of a men-optimal matching under the Mallows model with various \( \phi_m, \phi_w \) and \( n = 150 \). Confidence intervals are given in parentheses.
Algorithm performance in the Uniform model

Under the Uniform model, the outcomes of the DA algorithm have high sex-equality costs, as expected, since men- and women-optimal matchings have high sex-equality cost [58]. Other algorithms performed better: iBILS and PowerBalance showed the best performance and in most cases were able to find a sex-equal stable matching.

Fairness beyond the Mallows model

F.1 Other Correlated Models

We also analyzed the fairness of other correlated preference models known in the literature. Namely, we analyze the known results for some probabilistic models with hard-tier sets as well as those that consider correlation coefficients on preferences. It is unclear whether these models can be thought of as Mallows models (or Mallows mixture). Yet, similar to the Mallows model correlation intensity between the preferences of two sides might be a reasonable indicator in predicting the sex-equality cost of stable matchings.

Asymmetric correlation

The Discrete Uniform model from [71] is a probabilistic hard-tier preference model in which a subset of agents is always ranked higher than the rest of the set—referred to as hot and cold sets. Within a set, partners are sampled uniformly. According to experiments of [71] when men have uniform preferences on women, and women have discrete uniform preferences on men, the women-proposing DA has one of the lowest sex-equality cost and drastically outperforms many procedurally fair algorithms. In this case, women’s preferences are more correlated compared to men, so the women-optimal matching might become a sex-equal matching (or gives a good approximation of its sex-equality cost), coinciding with our findings for the Mallows model in asymmetric correlation markets.
Symmetric correlation

Celik and Knoblauch [16] proposes a correlation coefficient of preferences $\rho$. The coefficient computes the sum of agents’ average ranks in preferences of men (or women), normalized to take values between 0 (perfectly uncorrelated preferences) and 1 (identical preferences). The authors used the correlation coefficients of men $\rho_m$ and women $\rho_w$ as a predictor of men’s and women’s satisfaction by DA outcome. To generate instances with different correlation intensities they sampled from a hard-tier model while varying the number of tiers. They showed that a higher correlation in preferences of women sharply reduces their satisfaction by a men-optimal matching while making the other side slightly better off. These insights are consistent with our observations (e.g. in Figure 5 $S_W$-pessimal rises, and $S_M$-optimal moderately decreases when $\phi_w$ decreases).