Behavioral Stable Marriage Problems

Andrea Martin[1], Kristen Brent Venable[1,2], and Nicholas Mattei[3]

[1] University of West Florida, Pensacola FL 32514, USA
[2] Institute for Human and Machine Cognition, Pensacola FL 32502, USA
[3] Tulane University, New Orleans LA 70118

Abstract. The stable marriage problem (SMP) is a mathematical abstraction of two-sided matching markets with many practical applications including matching resident doctors to hospitals and students to schools. Several preference models have been considered in the context of SMPs including orders with ties, incomplete orders, and orders with uncertainty, but none have yet captured behavioral aspects of human decision making, e.g., contextual effects of choice. We introduce *Behavioral Stable Marriage Problems (BSMPs)*, bringing together the formalism of matching with cognitive models of decision making to account for multi-attribute, non-deterministic preferences and to study the impact of well known behavioral deviations from rationality on two core notions of SMPs: stability and fairness. We analyze the computational complexity of BSMPs and show that proposal-based approaches are affected by contextual effects. We then propose and evaluate novel ILP and local-search-based methods to efficiently find optimally stable and fair matchings for BSMPs.

Keywords: Stable Marriage, Psychological Decision-making

1 Introduction

The stable marriage problem (SMP) has a variety of applications in the context of two-sided markets, including matching doctors to hospitals and students to schools [24]. Typically, n men and n women express their preferences, via a strict total order, over the members of the other sex. Solving an SMP typically means finding a matching between men and women satisfying certain properties including *stability*, where no man and woman who are not married to each other would both prefer each other to their partners or to being single. Another desirable property is fairness, aiming at a balance between the satisfaction of the two groups [14]. A rich literature has been developed for SMPs [14], and many variants have been studied, including when there is uncertainty in the preferences [2] or where preferences are expressed according to multiple attributes [8].

We explore the connection between how people make choices, the process of matching, and the notions of stability and fairness. We assume that the preferences of each agent are encapsulated via a Multi-alternative Decision Field Theory (MDFT) model [23], that is, by a dynamic cognitive model of choice, capable of capturing behavioral aspects of human decision making. We choose this model as it has been shown to capture choice behavior accurately in human studies [6], it is designed to handle multiple options and attributes [23], and since it strikes a balance between the expressiveness of the underlying preference structure and its psychological underpinnings. One of the core characteristics of

MDFT is that choices may change based on the particular subset presented at any given point. This raises questions for classical matching algorithms, such as Gale-Shapley [10], a proposal based method where an agent is selecting alternatives to propose to from an increasingly smaller subset.

From an AI point of view, we extend the state of the art on SMPs by introducing the first framework that incorporates simultaneously multi-attribute preferences with uncertainty and cognitive modeling of bounded-rationality. From a cognitive science perspective, our work provides a psychologically grounded computational model of how humans may respond to matching procedures.

Our work is related to that in Aziz et al. [1] where the authors consider SMPs with uncertain pair-wise preferences, equivalent to considering the choice probabilities induced on subsets of size two by MDFT. While the considered notions of stability are closely related, MDFTs also induce choice probability distributions over subsets of any size, which play an important role for proposal-based methods. The different models of uncertainty in preferences considered in [2] are less closely related as they do not consider choice probability distributions over all subsets of members of the opposite group, as in MDFTs. Preferences expressed via multiple attributes have also been considered in the literature and, more recently, in [22] and [8]. However, in both cases preferences are qualitative, rather than quantitative as in our case. Fairness in matchings has received new attention recently and new algorithms for different definitions of fairness [9], procedural approaches to enforce sex equal stable matchings [27, 26, 11] and new preference models including bounded lists [20] have been proposed. However, none of these works focus on behavioral models of choice, as in this paper.

Contribution. We define Behavioral Stable Matching Problems (BSMP), where agents express preferences via MDFTs and analyze the computational complexity of several problems related to stability and fairness. We study the impact of behavioral effects on proposal based matching algorithms. We propose novel algorithms for finding maximally stable and fair stable matchings in BSMPs, which we analyse experimentally in terms of the efficiency, stability, and fairness of the returned matchings.

2 Multialternative Decision Field Theory (MDFT)

MDFT [4] models preferential choice as an accumulative process in which the decision maker attends to a specific attribute at each time point in order to derive comparisons among options, and update his preferences accordingly. Ultimately the accumulation of those preferences forms the decision maker's choice. In MDFT an agent is confronted with multiple options and equipped with an initial personal evaluation for them according to different criteria, called attributes. For example, a student who needs to choose a main course among those offered by the cafeteria will have in mind an initial evaluation of the options in terms of how tasty and healthy they look. More formally, MDFT, in its basic formulation [23], is composed of the following elements.

Personal Evaluation: Given a set of options $O = \{o_1, \ldots, o_k\}$ and set of attributes $A = \{a_1, \ldots, a_l\}$, the subjective value of option o_i on attribute a_j is denoted by m_{ij} and stored in matrix $\mathbf{M}$. In our example, let us assume that the cafeteria options are *Salad (S)*, *Burrito (B)* and *Vegetable pasta (V)*. Matrix

M, containing the student's preferences, could be defined as shown in Figure 1 (left), where rows correspond to the options (S, B, V) and the columns to the attributes $Taste$ and $Health$.

$$M = \begin{vmatrix} 1 & 5 \\ 5 & 1 \\ 2 & 3 \end{vmatrix} \quad C = \begin{vmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{vmatrix} \quad S = \begin{vmatrix} +0.9000 & 0.0000 & -0.0405 \\ 0.0000 & +0.9000 & -0.0047 \\ -0.0405 & -0.0047 & +0.9000 \end{vmatrix}$$

Fig. 1. Evaluation (M), Contrast (C) and Feedback (S) matrix.

Attention Weights: Attention weights express the attention allocated to each attribute at a particular time t during deliberation. We denote them by vector $\mathbf{W}(t)$, where $W_j(t)$ represents the attention to attribute a_j at time t. We adopt the common simplifying assumption that, at each point in time, the decision maker attends to only one attribute [23]. Thus, $W_j(t) \in \{0, 1\}$ and $\sum_j W_j(t) = 1$, $\forall t, j$. In our example, where we have two attributes, at any point in time t, we will have $\mathbf{W}(t) = [1, 0]$, or $\mathbf{W}(t) = [0, 1]$, representing that the student is attending to, respectively, $Taste$ or $Health$. The attention weights change across time according to a stationary stochastic process with probability distribution $\mathbf{p}$, where p_j is the probability of attending to attribute a_j. In our example, defining $p_1 = 0.55$ and $p_2 = 0.45$ means that at each point in time, the student will be attending $Taste$ with probability 0.55 and $Health$ with probability 0.45; i.e., $Taste$ matters slightly more than $Health$ to this student.

Contrast Matrix: Contrast matrix **C** is used to compute the advantage (or disadvantage) of an option with respect to the other options. In the MDFT literature [5, 23, 6], **C** is defined by contrasting the initial evaluation of one alternative against the average of the evaluations of the others, as shown for the case with three options in Figure 1 (center).

At any moment in time, each alternative in the choice set is associated with a **valence** value. The valence for option o_i at time t, denoted $v_i(t)$, represents its momentary advantage (or disadvantage) when compared with other options on some attribute under consideration. The valence vector for k options $o_1, \ldots, o_k$ at time t, denoted by column vector $\mathbf{V}(t) = [v_1(t), \ldots, v_k(t)]^T$, is formed by $\mathbf{V}(t) = \mathbf{C} \times \mathbf{M} \times \mathbf{W}(t)$. In our example, the valence vector at any time point in which $\mathbf{W}(t) = [1, 0]$, is $\mathbf{V}(t) = [1 - 7/2, 5 - 3/2, 2 - 6/2]^T$.

In MDFT, preferences for each option are accumulated across iterations of the deliberation process until a decision is made. This is done by using **Feedback Matrix S**, which defines how the accumulated preferences affect the preferences computed at the next iteration. This interaction depends on how similar the options are in terms of their initial evaluation expressed in **M**. Intuitively, the new preference of an option is affected positively and strongly by the preference it had accumulated so far, while it is inhibited by the preference of other options which are similar. This lateral inhibition decreases as the dissimilarity between options increases. Figure 1 (right) shows **S** computed for our running example following the MDFT standard method described in [16].

At any moment in time, the preference of each alternative is calculated by $\mathbf{P}(t + 1) = \mathbf{S} \times \mathbf{P}(t) + \mathbf{V}(t + 1)$, where $\mathbf{S} \times \mathbf{P}(t)$ is the contribution of the past preferences and $\mathbf{V}(t + 1)$ is the valence computed at that iteration. Starting

with $\mathbf{P}(0) = 0$, preferences are then accumulated for either a fixed number of iterations, and the option with the highest preference is selected, or until the preference of an option reaches a given threshold. In the first case, MDFT models decision making with a *specified* deliberation time, while, in the latter, it models cases where deliberation time is *unspecified* and choice is dictated by the accumulated preference magnitude.

Definition 1 (Multi-Alternative Decision Theory (MDFT) Model). *Given set of options $O = \{o_1, \ldots, o_k\}$ and set of attributes $A = \{a_1, \ldots, a_l\}$, an MDFT Model is defined by the n-tuple $Q = \langle \boldsymbol{M}, \boldsymbol{C}, \boldsymbol{p}, \boldsymbol{S} \rangle$, where: $\boldsymbol{M}$ is the $k \times l$ personal evaluation matrix; $\boldsymbol{C}$ is the $k \times k$ contrast matrix; $\boldsymbol{p}$ is a probability distribution over attention weights vectors; and $\boldsymbol{S}$ is the $k \times k$ feedback matrix.*

Moreover, we will denote with s-MDFT, resp. u-MDFT, models with specified, resp. unspecified, deliberation time. We will, however, omit such prefixes whenever the discussion applies to both types of models.

Different runs of the same MDFT model may return different choices due to the uncertainty on the attention weights distribution. The model can be run on a subset of options $Z \subseteq O$ of size $k' \leq k$, by eliminating from $\mathbf{M}$ all of the rows corresponding to options not in Z and resizing the contrast matrix and the feedback matrix to size k'. An MDFT induces a choice probability distribution over the options in a set. More formally:

Definition 2 (Choice probability distributions induced by an MDFT model). *Given an MDFT model $Q = \langle \boldsymbol{M}, \boldsymbol{C}, \boldsymbol{p}, \boldsymbol{S} \rangle$, defined over options set O and with attributes in A, we define the set of choice probability distributions $\{p_Z^Q | \forall Z, Z \subseteq O\}$, containing a probability distribution, denoted p_Z^Q, for each subset Z of O, where $p_Z^Q(z_i)$ is the probability that option $z_i \in Z$ is chosen when Q is run on subset of options Z.*

If we run the model a sufficient number of times on the same set, we obtain a proxy of its choice probability distribution. We note that these choice distributions may violate the regularity principle, which states that, when extra options are added to a set, the choice probability of each option can only decrease. This allows MDFT to effectively replicate bounded-rational behaviors observed in humans [5] such as the *similarity effect*, by which adding a new similar candidate decreases the probability of an option to be chosen, and the *compromise effect* where including a diametrically opposed option may increase the choice probability of a compromising one [23].

There is an interesting relation between the type of MDFT models and the stochastic transitivity of the induced probabilistic preference relation.

Definition 3 (Stochastic Transitivity). *Given MDFT model Q, defined over option set O, and induced choice probability distributions p_Z^Q, consider every $A, B, C \in O$ such that $p_{\{A,B\}}^Q(A) \geq 0.5$ and $p_{\{B,C\}}^Q(B) \geq 0.5$. If $p_{\{A,C\}}^Q(A) \geq 0.5$, then Weak Stochastic Transitivity (WST) holds. If $p_{\{A,C\}}^Q(A) \geq min\{p_{\{A,B\}}^Q(A), p_{\{B,C\}}^Q(B)\}$, then Moderate Stochastic Transitivity (MST) holds. If $p_{\{A,C\}}^Q(A) \geq max\{p_{\{A,B\}}^Q(A), p_{\{B,C\}}^Q(B)\}$, then Strong Stochastic Transitivity (SST) holds.*

Clearly SST implies MST and MST implies WST. In [7] it is shown that pairwise choice probabilities induced by s-MDFT models satisfy MST (and thus also WST), while only WST holds for those induced by u-MDFT models [4]. SST is not satisfied by MDFT models in general and, indeed, a systematic violation of SST by humans as been demonstrated by several behavioral experiments [21].

3 Stable Marriage Problems (SMPs)

In a *stable marriage problem* (SMP), we are given a set of n men $M = \{m_1, \ldots, m_n\}$, and a set of n women $W = \{w_1, \ldots, w_n\}$, where each strictly orders all members of the opposite gender. We wish to find a one-to-one matching s, of size n such that every man m_i and woman w_j is matched to some partner, and no two people of opposite sex who would both rather be married to each other than to their current partners; also called a *blocking* pair. A matching with no blocking pairs always exists and is said to be *stable* [19].

The Gale-Shapley Algorithm [10] is a well-known algorithm to solve an SMP. It involves a number of rounds where each un-engaged man "proposes" to his most-preferred woman to whom he has not yet proposed. Each woman must accept, if single, or choose between her current partner and the proposing man. GS returns a stable marriage in $O(n^2)$. Finding stable matching in variants of SMPs, such as with ties and incomplete lists, is, instead, NP-complete [19].

The pairing generated by GS with men proposing is male optimal, i.e., every man is paired with his highest ranked feasible partner, and female-pessimal [14]. Thus, it is desirable to require stable matchings to also be *fair*, for example, by minimizing the *sex equality cost (SEC)*: $SEC(s) = \mid \sum_{(m,w)\in s}(pr_m(w)) - \sum_{(m,w)\in s}(pr_w(m)) \mid$, where $pr_x(y)$ denotes the position of y in x's preference.

For example, if we consider the SMP of size 3 with men preferences defined as $m_1 : w_1 > w_2 > w_3$, $m_2 : w_2 > w_1 > w_3$, and $m_3 : w_3 > w_2 > w_1$ and women preferences $w_1 : m_1 > m_2 > m_3$, $w_2 : m_3 > m_1 > m_2$ and $w_3 : m_2 > m_1 > m_3$, we have two stable matchings, $s_m = \{(m_1, w_1), (m_2, w_2), (m_3, w_3)\}$ and $s_w = \{(w_1, m_1), (w_2, m_3), (w_3,\ m_2)\}$, that are, respectively, male and female optimal and have a SEC of, respectively, 4 and 3.

Finding a stable matching with minimum SEC is strongly NP-hard and approximation techniques have been proposed for example in [17]. Local search approaches have been used extensively in SMPs to tackle variants for which there are no polynomial stability and/or fairness algorithms [12, 19, 11].

4 Behavioral Stable Marriage Problems (BSMPs)

Given a set of n men and n women where each women w_i (resp. man m_i) expresses her (resp. his) preferences over the men (resp. women) via an MDFT model $Q_{w_i} = \langle \mathbf{M_{w_i}}, \mathbf{C_{w_i}}, \mathbf{p_{w_i}}, \mathbf{S_{w_i}} \rangle$ (resp. $Q_{m_i} = \langle \mathbf{M_{m_i}}, \mathbf{C_{m_i}}, \mathbf{p_{m_i}}, \mathbf{S_{m_i}} \rangle$). Since, as described in Section 2, we adopt the standard definitions for contrast and feedback matrices $\mathbf{C}$ and $\mathbf{S}$, we will omit them for clarity, in what follows.

Definition 4 (Behavioral Profile). *A Behavioral Profile is a collection of n men and n women, where the preferences of each man and woman, x_i, on the members of the opposite group are given by an MDFT model $Q_{x_i} = \langle \mathbf{M_{x_i}}, \mathbf{p_{x_i}} \rangle$.*

While each individual can, in principle, use different attributes to express their preferences, similarly to the MDFT literature, we will assume two attributes for all MDFTs. Thus, for each group member x_i, his/her model expresses a (numerical) personal evaluation of each member of the opposite group with respect to two attributes in $\mathbf{M_{x_i}}$, and the importance of each attribute, $\mathbf{p_{x_i}}$ (see an example in Figure 2). By running the MDFT models many times we can approximate the induced choice probabilities (Def. 2). For the profile in Fig. 2 we have $p^{Q_{m_1}}_{\{w_1,w_2\}}(w_1) = 0.485$, $p^{Q_{m_2}}_{\{w_1,w_2\}}(w_1) = 0.556$, $p^{Q_{w_1}}_{\{m_1,m_2\}}(m_1) = 0.495$, and $p^{Q_{w_2}}_{\{m_1,m_2\}}(m_1) = 0.562$.

$$M_{w_1} = \begin{bmatrix} A_1 & A_2 \\ 8 & 2 \\ 2 & 8 \end{bmatrix} M_{w_2} = \begin{bmatrix} A_1 & A_2 \\ 2 & 8 \\ 8 & 2 \end{bmatrix} M_{m_1} = \begin{bmatrix} A_1 & A_2 \\ 8 & 2 \\ 2 & 8 \end{bmatrix} M_{m_2} = \begin{bmatrix} A_1 & A_2 \\ 2 & 8 \\ 8 & 2 \end{bmatrix}$$

Fig. 2. A behavioral profile. Attention probability fixed to $p(A_1) = 0.55$.

As for SMPs, a *matching* is a one-to-one correspondence between men and women. However, the notion of blocking pair becomes probabilistic.

Definition 5 (β-blocking). *Let B be a behavioral profile, and s one of its matchings. Consider pair $(m, w) \notin s$ and let Q_m, Q_w, be the MDFT models of, respectively, m and w, and $s(m)$ and $s(w)$ be their respective partners in s. We say pair (m, w) is β-blocking if $\beta = p^{Q_m}_{\{w,s(m)\}}(w) \times p^{Q_w}_{\{m,s(w)\}}(m)$.*

In other words, we say that pair (m, w), unmatched in s, is β-blocking if β is the joint probability of m choosing w instead of $s(m)$ according to Q_m and of w choosing m instead of $s(w)$ according to Q_w. The higher the β, the higher the probability that m and w will break the current matching. For example, (m_1, w_2) is 0.29-blocking for matching $s = \{(m_1, w_1), (m_2, w_2)\}$ given the behavioral profile in Figure 2.

Definition 6 (α-B-stable matching). *Let B be a behavioral profile, and s one of its matchings. We say that s is α-behaviorally-stable (α-B-stable), if $((1-\beta_1)\times \ldots \times (1-\beta_h)) \leq \alpha$, and α is the minimum value for which this holds, where β_i is the blocking probability of pair π_i, $i \in \{1, \ldots, h\}$, un-matched in s, and h is the number of blocking pairs, that is, $h = n \times (n-1)$, if s has n pairs.*

Intuitively, a matching is α-B-stable if the probability that none of the unmatched pairs is blocking is smaller than or equal to α. We note that 1-B-stability corresponds to stability in the classical sense. Given the pair-wise probabilities described earlier, we see that matching $s = \{(m_1, w_1), (m_2, w_2)\}$ is 0.514-B-stable for the profile in Figure 2.

Definition 7 (Behavioral Stable Marriage Problem (BSMP)). *Given behavioral profile B, the corresponding Behavioral Stable Marriage Problem (BSMP) is that of finding an α-B-Stable matching with maximum α.*

With abuse of notation, we will use BSMP and behavioral profile, as well as marriage and matching, interchangeably in what follows. Moreover, we will write s-BSMP, resp. u-BSMP, to denote a BSMP where all agents express their preferences via s-MDFTs, resp. u-MDFTs.

Given model Q_m of man m, we define the probability that m's choices will follow a particular linear order as follows.

Definition 8 (Induced probability on linear orders). *Consider MDFT model Q defined on option set O. Let us consider linear order $\omega = \omega_1 > \cdots > \omega_k$, $\omega_i \in O$, defined over O. Then, the probability of ω given Q is: $p^Q(\omega) = p^Q_O(\omega_1) \times p^Q_{\{O-\{\omega_1\}\}}(\omega_2) \times \cdots \times p^Q_{\{\omega_{k-1},\omega_k\}}(\omega_{k-1})$.*

While one of several ways to obtain a linearization, the one in Def. 8 is particularly intuitive as the probability of a linear order is defined as the joint probability that the first element in the order will be chosen by the MDFT model among all of the options, the second element will be chosen among the remaining options, and so forth. We now define the expected position as follows.

Definition 9 (Expected position). *Consider BSMP B, man m and model Q^m. The expected position of w in $m's$ preferences is defined as: $E[pr_m(w)] = \sum_{\omega \in L(W)} p^{Q_m}(\omega) \times pr_\omega(w)$, where $L(W)$ is the set of linear orders over the set of women W, and $pr_\omega(w)$ is the position of woman w in linear order ω.*

We can now define the sex equality cost for BSMPs.

Definition 10 (Sex equality cost (SEC)). *Given BSMP B and matching s, the sex equality cost of s is: $SEC(s) = \mid \sum_{(m,w)\in s} E[(pr_m(w))] - \sum_{(m,w)\in s} E[(pr_w(m))] \mid$*

Clearly, the lower SEC the more fair the matching. Figure 3 provides two examples of BSMPs and SECs for matchings.

5 Complexity Results

In this section we study the complexity of several problems in the context of behavioral profiles. In particular, we reconsider some of the results presented in [1] in light of our setting. For all our results, we assume that we begin with the probabilities induced on all sets of size two by the agents' MDFTs. As noted in Section 2, s-MDFTs induce MST pairwise preferences, u-MDFTs induce WST pairwise preferences and, if we relax both the constraint on the specified deliberation time and neutral starting point, the induced probabilistic preferences may violate WST. While it is known that MDFTs are very successful in capturing choice distributions exhibited in humans, a theoretical analysis of their exact expressive power is still an open problem.

The problems we consider are the following: StabilityProbability: Given a BSMP B and a matching s, find $\alpha \in [0, 1]$ such that s is α-B-stable; ExistPossiblyStableMatching: Does there exist an α-B-Stable matching with $\alpha > 0$?; MatchingwithHighestProbability: Compute an α-B-stable matching with maximum α; MaximallyFairMatching: Find a matching s with minimum sex equality cost; MaximallyFairStableMatching: Find a matching with minimum SEC among those that are α-B-stable with maximum α.

We note that the complexity of obtaining induced probability distributions has been shown to be polynomial for s-MDFTs [7], where an analytical derivation

from the parameters of the model is described. For u-MDFTs we leverage the fact that we can approximate such pairwise probabilities by running the model a sufficient number of times. When the size of the options set is fixed at two, the amount of time necessary to obtain these approximations for all of the agents in a u-BSMP grows linearly with the number of agents. Following [1], we define the certainly preferred relation where for agent w, $b \succ_w^{cert} c$ if and only if she chooses b over c with probability 1. All proofs are omitted due to lack of space.

Theorem 1. *For BSMPs,* StabilityProbability *is polynomially solvable.*

This result derives directly from Theorem 1 in [1]. We note that the fact that pairwise probabilities in the context of MDFTs are defined in terms of choice distributions over subsets of size two, implies that they are independent. From this we derive the fact that the probabilities of each member of a blocking pair preferring the alternative options to their current match are also independent, this is also observed by [1].

Theorem 2. ExistPossiblyStableMatching *is NP-complete even if one side of the market has linear preferences and the other side has weakly stochastic transitive (WST) pairwise probabilities.*

This result strengthens the statement of Theorem 2 in [1] by further restricting the preferences of one side of the market.

Lemma 1. *For s-BSMPs, an α-B-stable matching with $\alpha > 0$ always exists and can be found in polynomial time.*

This results is derived by linearizing the probabilistic preferences induced by the s-MDFTs in a specific way so to obtain and SMP the stable matchings of which are α-B-stable with $a > 0$ in the s-BSMP. An immediate consequence is:

Theorem 3. *For s-BSMPs,* ExistPossiblyStableMatching *is polynomially solvable. (See proof in Appendix)*

We now consider the complexity of MatchingwithHighestProbability.

Theorem 4. MatchingWithHighestStabilityProbability
is NP-hard, even if the certainly preferred relation is transitive for one side of the market and the other side has WST preferences.

In [1] it is shown that this problem is NP-hard even if the certainly preferred relation is transitive for one side of the market and the other side has deterministic linear orders. Our result for WST preferences is orthogonal, as WST does not imply transitive certainly preferred relation and vice-versa.

The complexity of this problem when one side has MST preferences remains an open problem. We conjecture NP-hardness remains as MST preferences are a subset of those where the certainly stable relation is transitive. We note that from Theorem 4 we can also immediately derive that MaximallyFairStableMatching is also NP-hard.

We conclude elaborating on MaximallyFairMatching. Let us denote with $F(n)$ the time required to run the MDFT model on a set of options of size n.

[4] The complexity of computing a linearization as described in Definition 8 is $O(nF(n))$. If we repeat this process a sufficiently large number of times, K, we can approximate the expected positions in $O(Kn^2F(n))$. Finding a maximally sex-equal, i.e., fair, stable matching is a well known NP-hard problem [17, 20]. The question of finding such a matching not subject to stability constraints remains an important open problem in the literature. As we will see in Section 6, we formulated an ILP to solve this problem to judge our algorithms effectiveness. Our results are summarized in Table 1.

Table 1. Complexity results. Problem names are abbreviated.

	StabProb	ExistPossStabMatch	MatchHighProb	MaxFairMatch	MaxFairStabMatch
WST	P	NP-complete	NP-hard	?	NP-hard
MST	P	P	?	?	?

6 Algorithms for BSMPs

In this section we outline several algorithms that find matchings with different properties. In particular we introduce two variants of Gale Shapley (B-GS and EB-GS), two integer linear program (ILP) formulations and two local search approaches. The details of the first three algorithms can be found in the Appendix.

Gale Shapley for BSMPs: B-GS and EB-GS. The Gale Shapley procedure can be extended in a straightforward way to BSMPs by invoking the relevant MDFT models when a proposal or an acceptance has to be made. We call this variant of GS, Behavioral Gale Shapley, denoted with B-GS. B-GS still converges, since the sets of available candidates shrink by one every time a proposal is made, but it is no longer deterministic and may return different matchings as a consequence of the non-determinism of the underlying MDFT models. We also define another variant of GS, that we call Expected Behavioral Gale Shapley (EB-GS), which runs GS on the SMP obtained considering the linear orders corresponding to expected positions (see Definition 9).

Algorithm FB-ILP. We developed an integer linear program (ILP) to find the most fair solution according to the SEC with no guarantees on stability. For each combination of man $m_i \in M$ and woman $w_j \in W$, $|M| = |W| = n$, we introduce a binary variable m_iw_j that takes value 1 if m_i is matched with w_j and 0 otherwise. The FB-ILP formulation also includes two $n \times n$ matrices (pos_M and pos_W) modeling expected positions of respectively women and men in each others preferences. The solution with the lowest SEC is then obtained by minimizing $SEC = |\sum_{i,j \in n} pos_M[i,j] \cdot m_iw_j - \sum_{i,j \in n} pos_W[j,i] \cdot m_iw_j|$, leveraging an indicator variables approach [3] to bypass the non-linearity.

Algorithm B-ILP. To find the optimal α-B-Stable solution with B-ILP, we begin with the same setup of FB-ILP. In addition, the B-ILP formulation uses an $n \times n$ matrix Pr_{m_i} where entry $Pr_{m_i}[j,k]$ gives the probability that man m_i prefers w_j to w_k. This matrix can be computed by running the BSMP of man m_i a sufficiently large number of times. Then, to address the fact that the product of the probabilities is a convex not linear function, and stability is a pairwise notion over a given matching, we introduce $\forall((i,j),(k,l)) \in \binom{\binom{n}{2}}{2}$ possible combinations

[4] As in the MDFT literature, we can assume constant number of attributes and assume $F(n)$ polynomial in n for both halting modes.

of pairs of pairs, an indicator variable $m_i w_j + m_k w_l$ to indicate that both $m_i w_j$ is matched and $m_k w_l$ is also matched. This allows us to compute the blocking probability of m_i and w_l as well as of m_k and w_j. Hence for every pair of possible marriages $m_i w_j + m_k w_l$ we can compute the probability that these four individuals are not involved in blocking pairs by taking the likelihood that they swap partners, formally let $block[(ij),(kl)] = (1 - Pr_{m_i}[l,j] * Pr_{w_l}[i,k]) * (1 - Pr_{m_k}[j,l] * Pr_{w_j}[k,i])$. To handle the convex constraint we simply take the log of this quantity and maximize using an indicator variable which we implement using the Gurobi *And* constraint.

The B-LS algorithm. B-LS, is a local search approach [15] that explores the space of matchings to find one with maximum α-B-stability starting from a randomly generated one. Each matching s is evaluated by its level α of behavioral stability. When we find a matching, we compute for each non-matched pair its β-blocking level. The neighborhood of a matching s consists of all the matchings that can be obtained from s by rotating a blocking pair (i.e, swapping partners) and is explored in decreasing order of β until a matching with a higher α-B-stability is found or the neighborhood is exhausted and search restarts from a randomly generated matching. The search ends after a max number of iterations, returning the matching with maximum α found so far.

Algorithm FB-LS Algorithm FB-LS is another local search approach designed to take in input a value α and return a matching with the lowest SEC that is also α-B-stable. Intuitively, FB-LS runs B-LS on the space of matchings meeting a certain level of fairness. We first run B-LS to compute the maximum level of α-B-stability achievable, denoted α_{max}. We also compute the SEC for the matching returned by this run of B-LS, called $se_{\alpha_{max}}$. We then fix the lowest level of behavioral stability that we consider reasonable, denoted α_{min}, with $\alpha_{min} \leq \alpha_{max}$. Then FB-LS performs an incremental search where for each SEC value, se, it launches B-LS to find the matching with maximum α-B-stability value, say α_{se} and with SEC cost se. FB-LS starts with $se = se_{\alpha_{max}}$ and decreases se until it no longer finds a matching with stability $\alpha_{se} \geq \alpha_{min}$.

7 Experimental Results

We first exemplify how contextual effects impact the α-B-stability of a matching returned by a proposal-based approach. The key point is that MDFT captures and replicates preference reversals that humans exhibit when options are added or deleted to a choice set. Thus, what may have emerged like a good choice among several options at proposal time, may not be dominating when only choice sets of size two are considered for stability. As seen in Figure 3 (a), on an instance of the compromise effect both B-GS and EB-GS return a matching which is sub-optimal w.r.t. α-B-stability with high probability. An analogous situation can be observed for the instance of the similarity effect shown in Figure 3(b). These examples show that, in general, there is no guarantee that a matching returned by B-GS or EB-GS will be optimal w.r.t. α-B stability. In the second column of the tables in Figures 3 we show the SEC of the matchings. Not surprisingly, there is no guarantee on the fairness nor, most importantly, on the "unfairness" (as instead is the case for GS for SMPs) of the returned matching, the latter being an effect of the non-deterministic behavioral models.

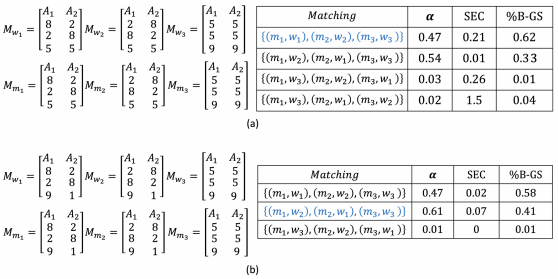

Fig. 3. Compromise (a) and Similarity effect (b), impact on GS. Profile (left) and results (right), for α-B stability value (α), Sex Equality Cost (SEC) and % of times returned by B-GS (%B-GS) out of 100 runs. EB-GS result in blue.

To test our algorithms in terms of efficiency and quality of the solutions we first generate 100 random BSMPs for each size n between 10 and 16 where the $\mathbf{M}$ matrices are of size $n \times 2$ and contain random preferences between 0 and 9. Attention weights probabilities are fixed to $p([0,1]) = 0.45$ and $p([1,0]) = 0.55$.

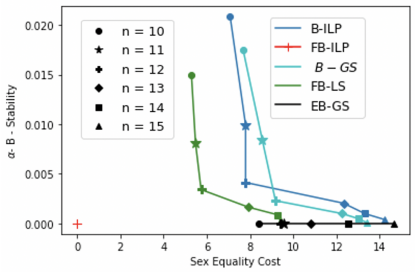

Fig. 4. Average α-B-Stability (y-axis) and SEC (x-axis) varying the number of agents.

Figure 4 shows α-B-Stability and SEC values of matchings returned by the algorithms averaged over the 100 instances. Each point on the lines represents the size of the problems from $n = 10$ to $n = 15$ moving from left to right. For $n = 16$ the ILP formulations timed-out at 6 hours while B-LS converges at around 340s (see Table 4). Not surprisingly, the quality of the solutions deteriorates as we move to larger problem sizes. The average results for B-ILP (dark blue-line) represent the optimal values for α-B-stability but exhibit average high SEC. In contrast, we can see how FB-LS (green line) allows to find matchings which have low SEC and are at most 30% less stable than optimal. As predicted, B-GS on average performs very poorly. At the bottom left corner we see the FB-ILP (red line) collapsed to a single point, as it always returns extremely unstable matchings of almost zero SEC. Our results showed very small variance in terms of α-B-stability, except for B-GS and EB-GS (see Table 3). Table 2 shows instead the SEC results with their standard deviations. All algorithms (except FB-ILP not shown since $\mu \cong 0$ and $\sigma^2 \cong 0$) have significant variance in terms of SEC, likely explained by the difference in preferences across instances. As expected, FB-LS exhibits the lowest SEC variance.

Table 2. Sex Equality Cost

	B-ILP		FB-LS		B-GS $max(\alpha)$		EB-GS	
# Agents	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
10	7.1	32.7	5.3	25.3	7.7	35.9	8.4	37.0
11	7.8	29.6	5.5	22.5	8.6	34.2	9.5	53.4
12	8.1	46.2	5.7	31.2	9.3	57.9	9.4	53.5
13	12.3	76.1	7.9	59.0	12.2	81.5	10.8	80.5
14	13.3	73.1	9.3	58.9	13.0	79.2	12.5	77.6
15	14.2	116.80	9.4	84.0	13.4	107.8	14.6	115.6

Table 3. α-B-Stability

	B-ILP		FB-LS		B-GS		EB-GS	
# Agents	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
10	0.0208	$5*10^{-4}$	0.0149	$3*10^{-4}$	0.0175	$5*10^{-4}$	$2*10^{-14}$	$6*10^{-26}$
11	0.0099	$2*10^{-4}$	0.0081	$2*10^{-4}$	0.0083	$2*10^{-4}$	$1*10^{-15}$	$1*10^{-28}$
12	0.0041	$3*10^{-5}$	0.0034	$3*10^{-5}$	0.0023	$2*10^{-5}$	$2*10^{-17}$	$5*10^{-32}$
13	0.0020	$5*10^{-6}$	0.0016	$4*10^{-6}$	0.0009	$2*10^{-6}$	$5*10^{-29}$	$2*10^{-55}$
14	0.0009	$3*10^{-6}$	0.0008	$2*10^{-6}$	0.0005	$2*10^{-6}$	$8*10^{-50}$	$4*10^{-97}$
15	0.0004	$3*10^{-7}$	0.0002	$2*10^{-7}$	0.0001	$5*10^{-8}$	$3*10^{-50}$	$5*10^{-98}$

Table 4. Average execution time for B-ILP, B-LS and B-GS varying n.

Algorithm	10	11	12	13	14	15	16
B-ILP	1.03s	2.74s	3.90s	6.61s	12.6s2	27.05s	N/A
B-LS	0.66s	2.01s	4.40s	15.17s	20.93s	24.94s	342s
FB-ILP	0.13s	0.15s	0.18s	0.22s	0.12s	0.12s	0.24s
FB-LS	2.83s	8.81s	35.16s	72.0s	90.223s	120.76s	941s
B-GS	1.93s	2.81s	3.18s	4.04s	4.55s	5.87s	7.2s
EB-GS	0.01s	0.015s	0.017s	0.02s	0.022	0.26	0.028s

The B-GS time is the average over 100 runs on the same instance. While B-GS and EB-GS are significantly faster, for each n they returned a maximally behaviorally stable matching only around 30% of the time. B-ILP and B-LS have comparable running times up to $n = 16$, where B-ILP doesn't terminate. It should also be noted that B-ILP, when terminating, always returns a maximally B-stable matching while B-LS does so around 88% of the time.

We also performed a convergence analysis on B-LS for $n = 16$ which showed B-LS plateaus at 300 iterations, corresponding to approximately 340s. We then tested B-LS on larger instances, generated under similar conditions, for $n \in \{20, 30, 40, 50\}$. Convergence was observed at, respectively, 500, 800, 1200 and 1900 iterations and average running times over 10 instance ranged from 896s for instances of size 20 to 146433s for size 50. We also note that, on average, the pre-processing times to compute the pairwise choice probabilities and the expected positions ranged between 16s for size $n = 10$ to 1592.5s for $n = 50$.

Our experimental results show that when the goal is to find a maximally stable matching, B-ILP is a viable and complete option for smaller problems. If fairness is also considered, then, FB-LS produces high quality solutions compromising between the two criteria while scaling reasonably well. This experimental study has also confirmed the negative impact of the underlying behavioral models on the quality of solutions returned by proposal based approaches.

8 Future work

We plan to consider the impact of behavioral models in one-to-many and many-to-many matching problems and their integration with other algorithms such as the Boston Mechanism [18]. We also plan to study methods proposed to achieve fairness over time which ties particularly well with the concept of repeated choices underlying the MDFT models [25].

Acknowledgements

Nicholas Mattei was supported by NSF Award IIS-2007955 and an IBM Faculty Research Award. K. Brent Venable are supported by NSF Award IIS-2008011.

References

1. Aziz, H., Biró, P., Fleiner, T., Gaspers, S., de Haan, R., Mattei, N., Rastegari, B.: Stable matching with uncertain pairwise preferences. In: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017. pp. 344–352. ACM (2017)
2. Aziz, H., Biró, P., Gaspers, S., de Haan, R., Mattei, N., Rastegari, B.: Stable matching with uncertain linear preferences. Algorithmica **82**(5), 1410–1433 (2020)
3. Bertsimas, D., Tsitsiklis, J.N.: Introduction to Linear Optimization, vol. 6. Athena Scientific Belmont, MA (1997)
4. Busemeyer, J.R., Diederich, A.: Survey of decision field theory. Mathematical Social Sciences **43**(3), 345–370 (2002)
5. Busemeyer, J.R., Townsend, J.T.: Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychological review **100**(3), 432 (1993)
6. Busemeyer, J., Gluth, S., Rieskamp, J., Turner, B.: Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. Trends Cogn Sci. **23**(3), 251–263 (2019)
7. Busemeyer, J., Townsend, J.: Fundamental derivations from decision field theory. Mathematical Social Sciences **23**(3), 255–282 (1992)
8. Chen, J., Niedermeier, R., Skowron, P.: Stable marriage with multi-modal preferences. In: Proceedings of the 2018 ACM Conference on Economics and Computation (ACM:EC). pp. 269–286. ACM (2018)
9. Cooper, F., Manlove, D.: Algorithms for new types of fair stable matchings. arXiv preprint arXiv:2001.10875 (2020)
10. Gale, D., Shapley, L.S.: College admissions and the stability of marriage. Amer. Math. Monthly **69**, 9–14 (1962)
11. Gelain, M., Pini, M., Rossi, F., Venable, K., Walsh, T.: Procedural fairness in stable marriage problems. In: 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011). pp. 1209–1210. IFAAMAS (2011)
12. Gelain, M., Pini, M.S., Rossi, F., Venable, K.B., Walsh, T.: Local search approaches in stable matching problems. Algorithms **6**(4), 591–617 (2013)
13. Gurobi Optimization, L.: Gurobi optimizer reference manual (2020), http://www.gurobi.com
14. Gusfield, D., Irving, R.W.: The Stable Marriage Problem: Structure and Algorithms. MIT Press, Cambridge, MA, USA (1989)
15. Hentenryck, P.V., Michel, L.: Constraint-based local search. MIT Press (2005)
16. Hotaling, J.M., Busemeyer, J.R., Li, J.: Theoretical developments in decision field theory: Comment on tsetsos, usher, and chater (2010). Psychological review (2010)
17. Iwama, K., Miyazaki, S., Yanagisawa, H.: Approximation algorithms for the sex-equal stable marriage problem. ACM Trans. Algorithms **7**(1), 2:1–2:17 (2010)
18. Kojima, F., Unver, M.: he "boston" school-choice mechanism: an axiomatic approach. Econ Theory **55**, 515–544 (2014)
19. Manlove, D.F.: Algorithmics of Matching Under Preferences, Series on Theoretical Computer Science, vol. 2. WorldScientific (2013)
20. McDermid, E., Irving, R.W.: Sex-equal stable matchings: Complexity and exact algorithms. Algorithmica **68**(3), 545–570 (2014)

21. Mellers, B., Biagini, K.: Similarity and choice. Psychological Review **101**, 505–518 (1994)
22. Miyazaki, S., Okamoto, K.: Jointly stable matchings. J Comb Optim **38**, 646–665 (2019)
23. Roe, R., Busemeyer, J., Townsend, J.: Multi-alternative decision field theory: A dynamic connectionist model of decision-making. Psychological Review **108**, 370–392 (2001)
24. Roth, A.E.: Who Gets What – and Why: The New Economics of Matchmaking and Market Design. Houghton Mifflin Harcourt (2015)
25. Sühr, T., Biega, A., Zehlike, M., Gummadi, K., Chakraborty, A.: Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019. pp. 3082–3092. ACM (2019)
26. Tziavelis, N., Giannakopoulos, I., Johansen, R.Q., Doka, K., Koziris, N., Karras, P.: Fair procedures for fair stable marriage outcomes. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020,. pp. 7269–7276. AAAI Press (2020)
27. Tziavelis, N., Giannakopoulos, I., Doka, K., Koziris, N., Karras, P.: Equitable stable matchings in quadratic time. In: Advances in Neural Information Processing Systems. pp. 457–467 (2019)

Appendix

A Proofs of Theorems

Theorem 1. *For BSMPs,* StabilityProbability *is polynomially solvable.*

Proof. This result derives directly from Theorem 1 in [1]. Given the pairwise probabilities induced by the BSMPs, we can compute the probability that a unmatched pair is not blocking in constant time. We then take the product over such pairs which are quadratic in number.

Theorem 2. ExistPossiblyStableMatching *is NP-complete even if one side of the market has linear preferences and the other side has weakly stochastic transitive (WST) pairwise probabilities.*

Proof. This results strengthens the statement of Theorem 2 in [1] by further restricting the preferences of one side of the market. There the authors reduce from ExistCompleteStableMatching in Stable Matching with Ties and Incompleteness (SMTIs) [19] to ExistPossiblyStableMatching when men have linear preferences and by leveraging the ability to define a cycle of length three of certainly preferred relations in the women's preferences. WST pairwise probabilities do not allow for cycles of length three comprised of certainly preferred relations. However, they do allow for cycles of length four as the one shown in Figure 5. This observation allows the proof to proceed in a very similar way as that of Theorem 2 in [1].

For the reader's convenience, we provide the complete proof below incorporating the extended cycle and associated modifications.

Given Theorem 1, we know that computing StabilityProbability for BSMPs is polynomially solvable. This implies that checking if a matching has a non-zero probability of being stable can be done in polynomial time, and thus the problem is in NP.

To prove NP-hardness, we follow the proof of Theorem 2 in [1] and we reduce from the problem of deciding whether an instance of SMTI admits a complete stable matching. This problem was shown to be NP-complete even if the ties appear only on the women's side, and each woman's preference list is either strictly ordered or consists entirely of a tie of size two [19].

Let $M = \{m_1, m_2, \ldots, m_n\}$ and $W = \{w_1, w_2, \ldots, w_n\}$ be the set of men and women in SMTI I. We create an instance of the pairwise probability model I' where women's preferences are WST as follows. We add 4 men and women: $m_{n+1}, m_{n+2}, m_{n+3}$ and m_{n+4} and $w_{n+1}, w_{n+2}, w_{n+3}$ and w_{n+4}. As in [1] we call acceptable partners in I proper partners in I'. For each man m_i, $i \in \{1, ..., n\}$, in the original instance I, we extend his strict preference ordering on his proper partners arbitrarily, by appending the four new women and his unacceptable partners in I in some arbitrary order. For every woman w_i, $i \in \{1, ...n\}$, in I, we create the pairwise preferences as follows. Firstly, w_i prefers every proper partner of hers to every new or unacceptable man. Secondly, w_i prefers each of the 4 new men to unacceptable men in I.

The pairwise preferences of w_i over her proper partners are defined in the same way as in [1]: w_i certainly prefers m_k to m_l if w_i strictly prefers m_k to m_l in I, and if w_i is indifferent between m_k and m_l in I then the corresponding pairwise probability is 0.5 in I'.

We then define the pairwise preferences of w_i over the 4 new men as in Figure 5. More in detail, w_i certianly prefers m_{n+1} to m_{n+2}, m_{n+2} to m_{n+3}, and m_{n+3} to m_{n+4} and m_{n+4} to m_{n+1}, while she is indifferent between m_{n+1} and m_{n+3}, and m_{n+2} and m_{n+4}. We note that these preferences form a cycle of length 4 and respect WST.

The preferences of w_i over the unacceptable original candidates are arbitrary. Similarly to [1], we let each of the four new men have all the original women at the top of his preference list ordered according to their indices, followed with new women w_{n+1}, w_{n+2}, w_{n+3} and w_{n+4} (in this order). Moreover, the four new women have m_{n+1}, m_{n+2}, m_{n+3} and m_{n+4} at the top of their strict preference lists, followed by the original men in an arbitrary order. (Note that every complete linear order implies pairwise probability preferences and satisfies WST). At this point we can show that there exists a complete weakly stable matching in I if and only if there is a matching with positive stability probability in I' following the exact same reasoning as in [1]. To see the first direction, let μ be a complete weakly stable matching in I. It is easy to see that if we extend μ with pairs (m_{n+1}, w_{n+1}), (m_{n+2}, w_{n+2}), (m_{n+3}, w_{n+3}) and (m_{n+4}, w_{n+4}) then the resulting matching μ' has positive probability of being stable in I' . This is because there is no pair which would be certainly blocking for μ'. Conversely, suppose that μ' is a complete matching in I' with positive probability of being stable (i.e., it has no certainly blocking pair). It can be shown that every original woman has to be matched with a proper partner. Suppose for a contradiction that w_i is the woman with the smallest index who is not matched to a proper partner. If w_i is matched to an original man who was unacceptable to her in I then w_i would form a certainly blocking pair with any of the four new men. In fact, note that w_i certainly prefers either of the four new men to her partner. Moreover, as none of the new men are matched to a original woman with index smaller than i, hence they all certainly prefer w_i to their partners. Suppose now that w_i is matched to one of the four new men. Then w_i would form a certainly blocking pair with the subsequent new man according to her cyclical preference. (For instance, if w_i is matched to m_{n+1} in μ' then she forms a certainly blocking pair with m_{n+4}.) This is because the subsequent new man cannot have any better partner, since all the women with smaller indices than i are matched to a proper partner. So we arrive at the conclusion that every original woman is matched with a proper partner. Since μ' does not admit a certainly blocking pair and all original women are matched with proper partners, the restriction of μ' to the original agents is a stable and complete matching in I.

Theorem 3. *For s-BSMPs,* ExistPossiblyStableMatching *is polynomially solvable.*

Proof. Consider BSMP B where all agents have s-MDFTs. For each man and woman, we extract a linear order from the pairwise probabilities induced by their s-MDFT thus obtaining an stable matching problem I. We then show that

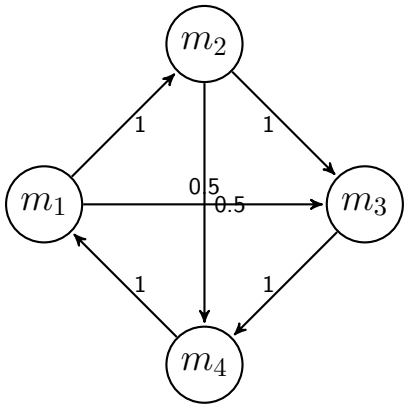

Fig. 5. Example of WST preferences with a cycle of length four. Directed edge represents dominance of source node on target node and the edges are annotated with the probabilities.

a matching is stable in I if and only if it is α-B-stable with $\alpha > 0$ in B. We illustrate the linearization for man m_i denoting with Q_i his s-MDFT model.

1. For every pair such that $P^{Q_i}_{\{w_k,w_j\}}(w_k) > 0.5$ we set $w_k >_{m_i} w_j$ in I. Note that, since the pairwise probabilities induced by an s-MDFt are MST, by doing this we cannot create any cycles in $>_{m_i}$ in I.
2. We perform the transitive closure adding all induced order relations.
3. At this point the only pairs that may still be not ordered in m_i's preferences in I must be such that $P^{Q_i}_{\{w_k,w_j\}}(w_k) = 0.5$. We order such remaining pairs (for example lexicographically) and we proceed in this order to pick one pair, order it in a random way, and then perform transitive closure.

It is easy to see that MST ensures that at the end of this process we obtain a linear order. Moreover each linearization requires polynomial time since in the worst case it performs a transitive closure $O(n^2)$ for each pair linearized in step 3, that is $O(n^2)$ times. Let μ be a complete stable matching in I. We know one exists [19]. Let's assume that μ is 0-B-stable in B. Then it must have a certainly blocking pair (m, w), where m prefers w to $\mu(m)$ and w prefers m to $\mu(w)$ with probability of 1 in B. If $P^m_{\{w,\mu(m)\}}(w) = 1$ in B then $w >_m \mu(m)$ in I. If $P^w_{\{m,\mu(w)\}}(m) = 1$ in B then $m >_w \mu(w)$ in I. That is, (m, w) is also a blocking pair in I, thus μ cannot be stable in I. This is a contradiction.

Theorem 5. MatchingWithHighestStabilityProbability
is NP-hard, even if the certainly preferred relation is transitive for one side of the market and the other side has WST preferences.

Proof. In [1] it is shown that this problem is NP-hard even if the certainly preferred relation is transitive for one side of the market and the other side has linear orders. Our result for WST preferences is orthogonal, as WST does

not imply transitive certainly preferred relation and vice-versa. The proof is adaptation of the one of Theorem 3 in [1] and leverages the same cycle described in Figure 5. Indeed, replacing the cycles of length three with cycles of length four, as the one depicted in Figure 5, does not affect the reasoning described in the proof. For the reader's convenience we provide the full details below, clarifying why the key steps still hold.

Following a similar reasoning as in [1], we derive the result by modifiyng the proof of Theorem 2. Let SMTI I and pairwise probability model with WST preferences I' be defined as in Theorem 2. We denote a new instance of a pairwise probability model with WST preferences I'' as follows. Whenever some women have cyclic certainly preferred relations in I', we modify the probabilities in these pairwise comparisons by a small value ϵ. That is, whenever a woman w_i certainly prefers man m_k to man m_l within a cycle in I', we modify the probability of the relation to $1 - \epsilon$ in I''. For example, the probabilities of the perimeter edges in Figure 5 would be set to $1 - \epsilon$. We note that, given how I' is defined, this modification will not cause violations of WST. Thus, we have no certainly preferred relations in any cycle in I''. However, as in [1] when considering the matching with the highest stability probability in I'', we can still articulate our reasoning along two cases with respect to the original NP-complete problem for I. Let's first assume that we have a complete stable matching for I. In this case this matching, extended with the four new pairs in I', will have a probability of being stable at least $\frac{1}{2^n}$ in both I' and I''. This is because every woman who is indifferent between some men has at most one tie of length two in her preference list in I by definition, and so if this woman is matched to one of the men in her tie then only the other man in this tie may block, which happens with 0.5 probability. Note that this step is not affected by the fact that we are using cycles of length four involving only the new men. On the other hand, if there exists no complete stable matching for I then we know from the proof of Theorem 2 that there always existed a certain blocking pair in I'. This certain blocking pair will now have a probability of $1 - \epsilon$ to be blocking, implying that any matching in this case has less than ϵ probability of being stable. Therefore, if we choose ϵ to be $0 < \epsilon < \frac{1}{2^n}$ we can use an algorithm which solves MatchingWithHighestStabilityProbability to decide the existence of a complete stable matching for SMTI efficiently.

B Algorithms

Algorithms B-GS and EB-GS. As we mentioned in the paper, the Gale Shapley procedure can be extended in a straightforward way to BSMPs by invoking the relevant MDFT models when a proposal or an acceptance has to be made. When man m is proposing, model Q_m will be run to select the woman to propose to among the set of women to whom m has not proposed yet. In fact, an MDFT model can be run on any subset of options by simply removing irrelevant rows from the personal evaluation matrix and resizing the other matrices (contrast and feedback). Similarly, when woman w, currently matched with man $\sigma(w)$ receives a proposal from m, the choice will be picked by running Q_w on the set $\{m, \sigma(w)\}$. We call this variant of GS, Behavioral Gale Shapley, denoted with

B-GS. While it is clear that B-GS still converges, since the sets of available candidates shrink by one every time a proposal is made, it is no longer deterministic and may return different matchings when run on the same BSMP. This is, of course, a consequence of the non-determinism of the underlying MDFT models.

We can also define another variant of GS that we call Expected Behavioral Gale Shapley (EB-GS). We first note that, given a man, we can extract a linear order from the expected positions of the women according to his MDFT model (breaking ties if needed). EB-GS corresponds to running GS on the profile of linear orders obtained in this fashion.

Algorithm FB-ILP. For each combination of $m_i \in M$ and $w_j \in W$, $|M| = |W| = n$, we introduce a binary variable $m_i w_j$ that takes value 1 if m_i is matched with w_j and 0 otherwise. We assume that for FB-ILP we have access to an $n \times n$ matrix $pos_M[i, j]$ where entry i, j gives us the expected position of w_j in the ranking of m_i, and the same matrix is available for the women, denoted pos_W.

Recall that finding the solution with lowest sex equality cost requires minimizing $SEC = |\sum_{i,j \in n} pos_M[i, j] \cdot m_i w_j - \sum_{i,j \in n} pos_W[j, i] \cdot m_i w_j|$. We cannot implement this absolute value directly as the optimization objective in Gurobi [13] as it is non-linear due to the presence of the absolute value. Since the SECs are always ≥ 0 we can overcome this using a standard trick in ILPs using indicator variables [3]. The SEC objective can be viewed as adding up the total man cost and the total woman cost, so we add indicator variables $tmc \geq 0$ and $twc \geq 0$ and minimize the difference between these two quantities. Hence, our full FB-ILP can be written as follows.

min $ind, s.t.,$	
(1) $\sum_{j \in n} m_i w_j = 1$	$\forall i \in n$
(2) $\sum_{i \in n} m_i w_j = 1$	$\forall j \in n$
(3) $\sum_{i,j \in n} m_i w_j = n$	
(4) $\sum_{i,j \in n} pos_M[i, j] \cdot m_i w_j = tmc$	
(5) $\sum_{i,j \in n} pr_W[j, i] \cdot m_i w_j = twc$	
(6) $twc \geq 0$	
(7) $twc \geq 0$	
(8) $twc - tmc = ind$	

In the constraints above (1) and (2) ensures that every man m_i has exactly one match across all possible women and every woman w_j has one match across all possible men. The redundant constraint (3) ensures that we have exactly n matches, i.e., everyone is matched. Constraint (4) captures the total cost to the men by multiplying the expected position by the indicator variables for the matches. Likewise constraint (5) captures the total woman cost. Constraint (8) is necessary to ensure that Gurobi handles our absolute value constraint correctly. We know that both $tmc \geq 0$ and $twc \geq 0$ from constraints (6) and (7), hence when Gurobi uses the Simplex Algorithm to solve, it will set $tmc = ind$ and $twc = 0$ if $ind > 0$ and otherwise we will have $tmc = 0$ and $tmc = -ind$. In either case we have a bounded objective function and we can find a solution if one exists.

Algorithm B-ILP. To find the optimal α-B-Stable solution with B-ILP, we begin with the same setup. For each $m_i \in M$ and $w_j \in W$ we introduce a

binary variable $m_i w_j$ defined as above. In addition, for B-ILP we assume that for each man and each woman we are given an $n \times n$ matrix Pr_{m_i} where entry $Pr_{m_i}[j,k]$ gives the probability that man m_i prefers w_j to w_k. This matrix can be computed by running the BSMP of man m_i a sufficiently large number of times.

There are two interrelated complications with formulating this probabilistic matching problem as an ILP: first we need the product of the probabilities which is a convex not linear function, and, second, stability is a pairwise notion over a given matching. To deal with both of these issues we introduce $\forall((i,j),(k,l)) \in \binom{\binom{n}{2}}{2}$ possible combinations of pairs of pairs, an indicator variable $m_i w_j + m_k w_l$ to indicate that both $m_i w_j$ is matched and $m_k w_l$ is also matched. This allows us to compute the blocking probability of m_i and w_l as well as of m_k and w_j. Given the formulation in [2], we know that we want to maximize the probability that *no blocking pair exists*. Hence for every pair of possible marriages $m_i w_j + m_k w_l$ we can compute the probability that these four individuals are not involved in blocking pairs by taking the likelihood that they swap partners, formally let $block[(ij),(kl)] = (1 - Pr_{m_i}[l,j] * Pr_{w_l}[i,k]) * (1 - Pr_{m_k}[j,l] * Pr_{w_j}[k,i])$. To handle the convex constraint we simply take the log of this quantity and maximize using an indicator variable we which we implement using the Gurobi *And* constraint. We can write the full program as follows.

max	$\sum_{\forall((i,j),(k,l)) \in \binom{\binom{n}{2}}{2}} pair_{m_i w_j + m_k w_l} * log(block[(ij),(kl)]), s.t.,$	
(1)	$\sum_{j \in n} m_i w_j = 1$	$\forall i \in n$
(2)	$\sum_{i \in n} m_i w_j = 1$	$\forall j \in n$
(3)	$\sum_{i,j \in n} m_i w_j = n$	
(4)	$AND(m_i w_j, m_k w_l) = pair_{m_i w_j + m_k w_l}$	$\forall((i,j),(k,l)) \in \binom{\binom{n}{2}}{2}$

In the constraints above (1) and (2) ensures that every man m_i has exactly one match across all possible women and every woman w_j has one match across all possible men. The redundant constraint (3) ensures that we have exactly n matches, i.e., everyone is matched. Constraint (4) uses the Gurobi [13] AND constraint to set the value of $pair\_m_i w_j + m_k w_l$ to be 1 if and only if both $m_i w_j$ and $m_k w_l$ are both 1. This allows us to capture all possible pairs of man/woman pairs and maximize the probability that no blocking pair occurs.

C Convergence Analysis for B-LS

The convergence analysis performed for $n = 16$ is shown in Fig. 6. While we depict the results of for seven runs we performed a total of 50 runs. The results indicated that B-LS plateaus after 300 iterations, corresponding to approximately 340s. B-LS does so around 88% of the time and returns a matching $1.006 * 10^{-}6$ far from optimal otherwise.

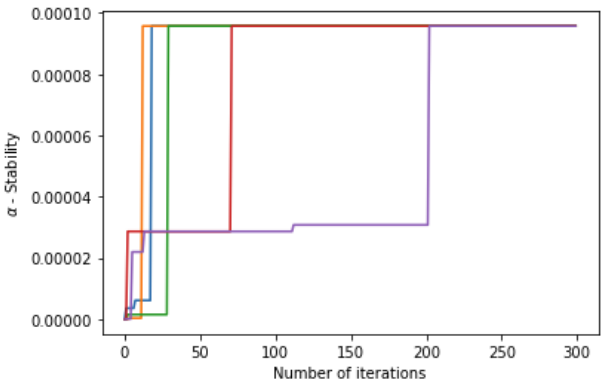

Fig. 6. Convergence of B-LS algorithm implementation with respect to α-B-Stability when $n = 16$.