Behavioral Stable Marriage Problems

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

1: University of West Florida, Pensacola, FL, USA
2: Institute for Human and Machine Cognition, Pensacola, FL, USA
3: Tulane University, New Orleans, LA, USA

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, incompleteness, and uncertainty. Yet, behavioral aspects of human decision making, including the similarity and compromise effects, which are captured by psychological choice models, have so far been neglected. We introduce *Behavioral Stable Marriage Problems (BSMPs)*, bringing together the formalism of matching with cognitive models of decision making to account for the impact of well known behavioral deviations from rationality on core notions of SMPs, such as, stability and fairness. We show that proposal-based approaches are affected by contextual effects and propose novel ILP and local-search-based methods to efficiently find optimally stable and fair matchings for BSMPs.

1 Introduction

The stable marriage problem (SMP) has a wide variety of applications, including matching doctors to hospitals, students to schools or, more generally, any two-sided market Roth [2015]. 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, such as, *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 Gusfield and Irving [1989]. A rich literature has been developed for SMPs Gusfield and Irving [1989], and many variants have been studied, including when there is uncertainty in the preferences Aziz et al. [2020] or where preferences are expressed according to multiple attributes Chen et al. [2018].

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 Roe et al. [2001a], that is, by a dynamic cognitive model of choice, capable of capturing behavioral aspects of human decision making. We choose this model for several reasons. MDFT belongs to a family of models based on the principle of *accumulation to threshold*, by which deliberation consists in a cumulative gathering of evidence until a certain threshold is reached. Among many proposed cognitive models, MDFT has been shown to capture choice behavior more accurately in human studies. Moreover, unlike other models, e.g. those proposed by Erev et al. [2017], MDFT is designed to handle scenarios with more than two options and where preferences are expressed in terms of multiple attributes. Other cognitive models, see for example Erev et al. [2017], Trueblood et al. [2014], rely heavily on strong psychological assumptions and directly incorporate behavioral observations in their implementation. In contrast, MDFT strikes a balance between the expressiveness of the underlying preference structure and its psychological underpinnings. In fact, within an MDFT, the initial evaluations of the options is expressed as an aggregation over features, and the MDFT models how this aggregation builds over time as separate components. Hence, an MDFT model is an appealing combination of cardinal preferences with psychological processes. This is an attractive feature from the point of view of integration with AI algorithms, and with matching procedures in particular. 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 Gale and Shapley [1962], 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, to best of our knowledge, 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 in the context of matching procedures. Integrating human preference models into the models of decision making in COMSOC and AI research more generally is an important direction for developing more applicable reserach Mattei [2020].

Contribution. We define a novel problem at the intersection of matching theory and cognitive theories of preferences: the Behavioral Stable Matching Problems (BSMP). This novel approach allows us to study the impact of behavioral effects and the MDFT choice model on proposal based matching algorithms. To account for this algorithmic integration of MDFT models into matching procedures, we propose two novel algorithms for finding maximally stable matchings, based on local search and ILP: an ILP method for finding fair matchings, and a local search method for finding matchings with maximal fairness for a specified threshold of stability. We validate our algorithms on an experimental evaluation of the proposed methods in terms of efficiency and of the stability and fairness of the returned matchings.

2 Multialternative Decision Field Theory (MDFT)

MDFT Busemeyer and Diederich [2002] models preferential choice as an accumulative process in which the decision maker attends to a specific attribute at each time 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 Roe et al. [2001b], is composed of the following elements.

Personal Evaluation: Given set of options $O = \{o_1, \dots, o_k\}$ and set of attributes $A = \{A_1, \dots, A_J\}$, 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 $\mathbf{M}$, 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 $Health$.

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

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

Attention Weights: Attention weights are used to express the attention allocated to each attribute at a particular time t during the deliberation. We denote them by vector

$\mathbf{W}(t)$ where $W_j(t)$ represents the attention to attribute j at time t. We adopt the common simplifying assumption that, at each point in time, the decision maker attends to only one attribute Roe et al. [2001a]. 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 $p_2 = 0.45$ would mean that at each point in time, the student will be attending $Taste$ with probability 0.55 and $Health$ with probability 0.45. In other words, $Taste$ matters slightly more than $Health$ to this particular student.

Contrast Matrix: Contrast matrix $\mathbf{C}$ is used to compute the advantage (or disadvantage) of an option with respect to the other options. In the MDFT literature Busemeyer and Townsend [1993], Roe et al. [2001a], Busemeyer et al. [2019], $\mathbf{C}$ 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 the 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 $\mathbf{M}$. Intuitively, the new preference of an option is affected positively and strongly by the preference it had accumulated so far, while it is strongly inhibited by the preference of other options which are similar. This lateral inhibition decreases as the dissimilarity between options increases. Figure 1 (right) shows $\mathbf{S}$ computed for our running example following the MDFT standard method described in Hotaling et al. [2010].

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.

Definition 1 (Multi-Alternative Decision Theory Model (MDFT Model))**.** *Given set of options $O = \{o_1, \ldots, o_k\}$ and set of attributes $A = \{A_1, \ldots, A_J\}$, 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 $k \times J$ 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.*

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'. If we run the model a sufficient number of times on the same set, we obtain a proxy of the choice probability distribution induced over the options in the set. More formally:

Definition 2 (Choice probability distribution 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.*

We note that the choice probability distributions induced by MDFT models 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 Busemeyer and Townsend [1993]. Consider an example involving an agent purchasing a car and considering the attributes of performance and fuel efficiency. Assume at first that there are two options, say, A, and B. Assume that A has better performance but poorer efficiency with respect to B. Behavioral studies have shown that introducing a third option C, similar to A, will decrease A's probability of being chosen and will increase B's probability of being selected instead Roe et al. [2001a]. This is known as the *similarity* effect. Now consider the case in which C is a compromising option with evaluations lying between those of A and B on both attributes. In this case human behavior will be skewed towards selecting C by the, so called, *compromise* effect Busemeyer and Diederich [2002].

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 person 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. Such a pair is called a *blocking* pair. In this setting, a matching with no blocking pairs always exists and is said to be *stable* Manlove [2013].

The Gale-Shapley Algorithm The *Gale-Shapley Algorithm* (GS) Gale and Shapley [1962] 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 (if she has one) and the proposing man. GS returns a stable marriage in $O(n^2)$.

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 Gusfield and Irving [1989]. 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.

Example 1. *Consider the following SMP of size 3.*

$$\begin{array}{ll} m_1 : w_1 > w_2 > w_3 & w_1 : m_1 > m_2 > m_3 \\ m_2 : w_2 > w_1 > w_3 & w_2 : m_3 > m_1 > m_2 \\ m_3 : w_3 > w_2 > w_1 & w_3 : m_2 > m_1 > m_3 \end{array}$$

Stable matchings $s_m = \{(m_1, w_1), (m_2, w_2), (m_3, w_3)\}$ and $s_w = \{(w_1, m_1), (w_2, m_3), (w_3, m_2)\}$ 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 Iwama et al. [2010]. Local search approaches have been used extensively in SMPs to tackle variants for which there are no polynomial stability and/or fairness algorithms Gelain et al. [2013], Manlove [2013], Gelain et al. [2011].

4 Related Work

The extension of classical SMPs that we present in this paper involves uncertainty in the preferences, multiple attributes, and bounded-rationality via a psychologically-grounded model of human behavior. While, to the best of our knowledge, this is the first attempt to handle all three of these aspects while maintaining cognitive plausibility, there is a rich literature on SMP extensions addressing one or more of them.

Aziz et al. [2017] consider SMPs with uncertain pair-wise preferences. From a knowledge representation point of view, the two frameworks are closely related. In fact, considering pairwise probabilities is equivalent to considering the choice probabilities induced on subsets of size two by MDFTs. However, a fundamental difference is that MDFTs also induce choice probability distributions over subsets of all other sizes. While this is irrelevant when focusing only on the notion of stability, it plays an important role when proposal-based methods are considered. The notion of α-behavioral stability, which we introduce in Section 5, coincides with that of possibly stable matching in Aziz et al. [2017] when $\alpha > 0$. The focus in Aziz et al. [2017] is on complexity results and indeed, their hardness results for finding maximally possibly stable matchings applies here. We concentrate on experimentally analyzing the behavior of different algorithmic approaches when preferences are represented via MDFT.

Different models of uncertainty in preferences are also considered in Aziz et al. [2020], where the complexity of different problem classes related to the probability of stability are explored. The closest model to ours among those studied in Aziz et al. [2020] is the lottery model, where each agent expresses his preferences over the opposite group as a probability distribution over linear orders. This is, however, different from the preference structure induced by MDFT models which consists of choice probability distributions over all subsets of members of the opposite group. The technical focus is also different, since we don't focus on complexity issues but rather on the interaction between realistic behavioral simulations and matching algorithms.

In our MDFT-based framework, the members of one group are evaluated quantitatively by each member of the opposite group according to multiple attributes. Preferences expressed via multiple attributes have been considered before in the literature and, more recently, in Miyazaki and Okamoto [2019] and Chen et al. [2018]. In both of these works the preferences are expressed qualitatively and consists of collections of linear orders. Moreover, the concepts of stability they define maintain the preference lists corresponding to different attributes separate. In contrast, in our setting the preferences according to different attributes are merged by the deliberation simulation into a choice or choice probabilities with the intent of replicating human behavior.

In this work we are concerned with matching that are both stable *and* fair. This area has received new attention recently with the growing conversation around fairness and equitability in AI systems Rossi and Mattei [2019], Lee et al. [2019], Loreggia et al. [2018]. This has resulted in recent works on new algorithms for different definitions of fairness Cooper and Manlove [2020b], besides the sex equality we consider here, including parameterized complexity of matchings with minimal egalitarian cost Gupta et al. [2019]. Other work has focused on the complexity of sex equal stable matchings through the use of fair procedures, where not all proposals happen on one side Tziavelis et al. [2019, 2020], Gelain et al. [2011] as well as complexity for preference models other than MDFT including bounded lists McDermid and Irving [2014] and in the general case Iwama et al. [2010]. While all these works focus on various forms of stability, equity, and preference model, none of them has investigated these concepts using a choice model as complex as MDFT and the resulting preference structures nor considered the behavioral aspects.

Finally, worth mentioning, is a recent work applying DFT (that is, MDFT for binary choices) to mimic the decision-making processes involving multiple agents Lee and Son [2016]. In particular, the authors present a version of DFT which accounts for forgetting

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

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

Figure 2: A behavioral profile. Attention weights probability fixed at $p(A_1) = 0.55$ and $p(A_2) = 0.45$.

to model multi-agent decision-making and the stability of a decision under the dynamics of opinion formation. While based on the same behavioral model, the framework described in Lee and Son [2016] is quite different from ours, as it considers a hierarchical network model of social choice rather than matching in two sided markets. Furthermore, the combination between DFT and AI techniques has been investigated in Martin and Venable [2018], where the authors designed a sequential procedure that uses DFT and soft constraints to model decision making over a set of interdependent choices. While using the same psychological model, the framework in Martin and Venable [2018] does not consider a multi-agent scenario and only explores the interplay between modeling choices using a DFT model and propagating their effects via constraint-based inference.

5 Behavioral Stable Marriage Problems (BSMPs)

In this section we formally define Behavioral Stable Marriage Problems. We are given a set of n men and n women. 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 the sake of clarity, in what follows.

Definition 3 (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 represented by an MDFT model $Q_{x_i} = \langle \mathbf{M_{x_i}}, \mathbf{p_{x_i}} \rangle$.*

We note that each individual can, in principle, use different attributes to express their preferences over the members of the other group. However, in all of our examples and experiments we assume two attributes. 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 $p^{Q_{w_2}}_{\{m_1,m_2\}}(m_1) = 0.562$.

As for SMPs, a *matching* is a one-to-one correspondence between men and women. However, in our setting, the answer to the question wether an individual would break his/her current matching and elope with another partner becomes probabilistic.

Definition 4 (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 equal to 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. As an example, pair (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 5 (α-B-stable matching). *Let B be a behavioral profile, and s one of its matchings. We say that s is α-behaviorally-stable (abbreviated, 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 or equal than α. We note that 1-B-stability corresponds to stability in the classical sense. The notions of β-blocking pair and α-B-stability require only choices over subsets of size 2 for which we can approximate the induced probabilities. 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. We conclude this section with the formal definition of Behavioral Stable Marriage Problem.

Definition 6 (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.

6 Fairness

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

Definition 7 (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: $p^Q(\omega) = p^Q_O(\omega_1) \times p^Q_{\{O-\{\omega_1\}\}}(\omega_2) \times \cdots \times p^Q_{\{\omega_{k-1},\omega_k\}}(\omega_{k-1})$.*

Intuitively, 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 by the MDFT model from the remaining options, and so forth. We now define the expected position as follows.

Definition 8 (Expected position). *Consider BMSP B, man m and model Q^m. The expected position of woman w in $m's$ preferences is defined as follows: $E[pr_m(w)] = \sum_{\omega \in L(W)} p^{Q_m}(\omega) \times pr_\omega(w)$, 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 9 (Sex Equality cost). *Given BSMP B and matching s we define the sex equality cost of s as: $SEC(s) = \mid \sum_{(m,w)\in s} E[(pr_m(w))] - \sum_{(m,w)\in s} E[(pr_w(m))] \mid$.*

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

Since computing the expected position of an option is computationally prohibitive, we obtain an approximation using the MDFT model to build linear orders. We do this by choosing a first element, then a second one from the remain set, and so on. By repeating this process a sufficiently large number of times, we can approximate the probability of linear orders given the MDFT and, thus, obtain an approximate value for the expected positions.

7 The Gale Shapley Algorithm and BSMPs

In this section we show how, on one hand, GS can be easily adapted to run on BSMPs while, on the other, behavioral effects may negatively impact the α-B stability of the matching.

Algorithms 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. 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. Similarly, when woman w, currently matched with, say, man m' receives a proposal from m, the choice will be picked by running Q_w on the set $\{m, m'\}$. 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.

Impact of Behavioral Effects. We now discuss how contextual effects impact the α-B-stability of a matching returned by a proposal-based approach.

Consider the compromise effect, modelling the tendency humans have to pick an option in the middle when confronted with others characterized by asymmetric strengths. An instance is shown in Figure 3(a) where we see that m_3 (resp. w_3) is the compromising option for women w_1 and w_2 (resp. for men m_1 and m_2), and is the preferred option for w_3 (resp. m_3). When proposals are made and all options are available, m_3 (resp. w_3) will be preferred by any woman (resp. man). However, for m_1, m_2, w_1 and w_2 every other choice between two alternatives is between incomparable options, yielding high uncertainty in the outcome of the MDFT model. As seen in Figure 3 (a), both B-GS and EB-GS return a matching which is sub-optimal w.r.t. α 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 sex equality cost of the matchings. Not surprisingly, there is no guarantee on the fairness of the matchings returned by B-GS or EB-GS. However, most importantly, there is also no guarantee on its "unfairness" (as instead is the case for GS on SMPs) which is again an effect of the non-determinism injected by the behavioral models.

$$M_{w_1} = \begin{bmatrix} A_1 & A_2 \\ 8 & 2 \\ 2 & 8 \\ 5 & 5 \end{bmatrix} M_{w_2} = \begin{bmatrix} A_1 & A_2 \\ 2 & 8 \\ 8 & 2 \\ 5 & 5 \end{bmatrix} M_{w_3} = \begin{bmatrix} A_1 & A_2 \\ 5 & 5 \\ 5 & 5 \\ 9 & 9 \end{bmatrix}$$

$$M_{m_1} = \begin{bmatrix} A_1 & A_2 \\ 8 & 2 \\ 2 & 8 \\ 5 & 5 \end{bmatrix} M_{m_2} = \begin{bmatrix} A_1 & A_2 \\ 2 & 8 \\ 8 & 2 \\ 5 & 5 \end{bmatrix} M_{m_3} = \begin{bmatrix} A_1 & A_2 \\ 5 & 5 \\ 5 & 5 \\ 9 & 9 \end{bmatrix}$$

Matching	$\boldsymbol{\alpha}$	SEC	%B-GS
$\{(m_1, w_1), (m_2, w_2), (m_3, w_3)\}$	0.47	0.21	0.62
$\{(m_1, w_2), (m_2, w_1), (m_3, w_3)\}$	0.54	0.01	0.36
$\{(m_1, w_3), (m_2, w_2), (m_3, w_1)\}$	0.03	0.26	0.01
$\{(m_1, w_3), (m_2, w_1), (m_3, w_2)\}$	0.02	1.5	0.04

(a)

$$M_{w_1} = \begin{bmatrix} A_1 & A_2 \\ 8 & 2 \\ 2 & 8 \\ 9 & 1 \end{bmatrix} M_{w_2} = \begin{bmatrix} A_1 & A_2 \\ 2 & 8 \\ 8 & 2 \\ 9 & 1 \end{bmatrix} M_{w_3} = \begin{bmatrix} A_1 & A_2 \\ 5 & 5 \\ 5 & 5 \\ 9 & 5 \end{bmatrix}$$

$$M_{m_1} = \begin{bmatrix} A_1 & A_2 \\ 8 & 2 \\ 2 & 8 \\ 9 & 1 \end{bmatrix} M_{m_2} = \begin{bmatrix} A_1 & A_2 \\ 2 & 8 \\ 8 & 2 \\ 9 & 1 \end{bmatrix} M_{m_3} = \begin{bmatrix} A_1 & A_2 \\ 5 & 5 \\ 5 & 5 \\ 9 & 5 \end{bmatrix}$$

Matching	$\boldsymbol{\alpha}$	SEC	%B-GS
$\{(m_1, w_1), (m_2, w_2), (m_3, w_3)\}$	0.47	0.02	0.58
$\{(m_1, w_2), (m_2, w_1), (m_3, w_3)\}$	0.61	0.07	0.41
$\{(m_1, w_3), (m_2, w_2), (m_3, w_1)\}$	0.01	0	0.01

(b)

Figure 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.

8 Integer Linear Programs for BSMPs

In order to test the efficacy of our algorithms we first developed integer linear program (ILP) formulations to find solutions that are maximally α-B-Stable, which we call B-ILP, as well as a formulation to find the most fair solution according to the sex equality cost with no guarantees on stability, FB-ILP. There is a long history of using ILP formulations for various versions of stable marriage Roth et al. [1993] and matching problems Lian et al. [2018] and even SAT encoding Drummond et al. [2015]. These are implementations for our simple baselines and optimizing them for deployment would be an interesting area of future work Pettersson et al. [2021].

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 Gurobi Optimization [2020] 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 Bertsimas and Tsitsiklis [1997]. 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-LP 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. 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 Aziz et al. [2020], 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 Gurobi Optimization [2020] 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.

9 Local search approaches for BSMPs

We present two algorithms based on local-search (LS) to find matchings with either high α-B-stability or with both a guaranteed level of α-B-stability and a low SEC. LS algorithms do not have any theoretical guarantee of returning optimal solutions, but often produce near-optimal solutions and scale better than complete procedures Hentenryck and Michel [2005].

The B-LS algorithm. B-LS, explores the space of matchings to find one with maximum α-B-stability. 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). B-LS explores the neighborhood by rotating blocking pairs in decreasing order of β until a matching with a higher α-B-stability is found or the neighborhood is exhausted. In the latter case, the 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 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. This is done by discarding any matching that does not meet the fairness requirement while exploring the neighborhood. In order to use FB-LS, we first run B-LS on the unconstrained space. This allows us to compute the maximum level of α-B-stability achievable, lets call it 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 a matching with maximum α-B-stability value, say α_{se} and with SEC cost se. FB-LS starts with $se = se_{\alpha_{max}}$ and gradually decreases se until it no longer finds a matching with behavioral stability $\alpha_{se} \geq \alpha_{min}$.

10 Experimental Results

We 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 $p([1,0]) = 0.55$.

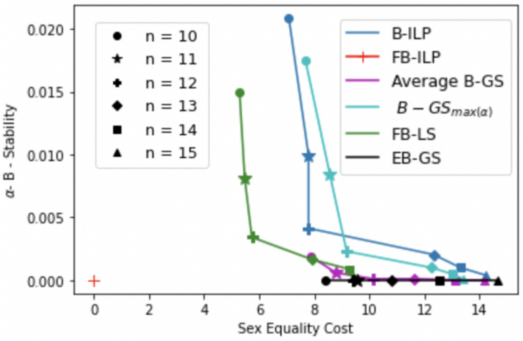

Figure 4: Average α-B-Stability (y-axis) and SEC (x-axis) when varying the number of agents.

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 2).

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 and remains sub-optimal even if the best values, instead of the average, are considered (B-GS$_{max(\alpha)}$). At the bottom left corner we see the FP-ILP (red line) collapsed to a single point, as it always returns extremely unstable matchings of almost zero SEC. Our results show very small variance in terms of α-B-stability, except for B-GS and EB-GS (see Table 3). Table 1 shows instead the SEC results plotted in Figure 4 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.

On average the pre-processing times to compute the pairwise choice probabilities and the expected positions ranged between 16s and 73s and 10.5s and 36s, respectively. In Table 2 we show the running time for all of the algorithms. The B-GS time corresponds to running the algorithm 100 times 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 $n = 16$ when the ILP method doesn't terminate. The convergence analysis performed for $n = 16$ is shown in Fig. 5. We can see that, on average, B-LS plateaus at 500 iterations, corresponding to approximately 340s. 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 and returns a matching $1.006 * 10^-6$ far from optimal otherwise. 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 well with the size of the problem. This experimental study has also confirmed the negative impact of the underlying behavioral models on the quality of solutions returned by proposal based approaches.

	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 1: Sex Equality Cost mean (μ) and standard deviation (σ^2).

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

Table 2: Average execution time for B-ILP, B-LS and B-GS when varying the number of agents.

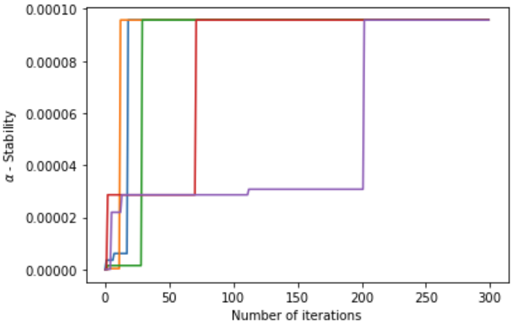

Figure 5: Convergence of B-LS algorithm implementation with respect to α-B-Stability when $n = 16$

	B-ILP		FB-LS		B-GS $max(\alpha)$		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 3: α-B-Stability

11 Future work

In the future we would like consider the impact of behavioral models in more complex settings, such one-to-many and many-to-many matching problems by studying their integration with other matching algorithms such as the Boston Mechanism Kojima and Unver [2014] and other applied matching mechanisms. We also plan to investigate further the interplay between fairness and behavioral modeling in algorithms targeting fairness both at the matching and the procedural level Tziavelis et al. [2020], Cooper and Manlove [2020a] and in methods proposed to achieve fairness over time which ties particularly well with the concept of repeated choices underlying the MDFT models Sühr et al. [2019].

References

Haris Aziz, Péter Biró, Tamás Fleiner, Serge Gaspers, Ronald de Haan, Nicholas Mattei, and Baharak Rastegari. Stable matching with uncertain pairwise preferences. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017*, pages 344–352. ACM, 2017.

Haris Aziz, Péter Biró, Serge Gaspers, Ronald de Haan, Nicholas Mattei, and Baharak Rastegari. Stable matching with uncertain linear preferences. *Algorithmica*, 82(5):1410–1433, 2020.

Dimitris Bertsimas and John N Tsitsiklis. *Introduction to Linear Optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

Jerome R Busemeyer and Adele Diederich. Survey of decision field theory. *Mathematical Social Sciences*, 43(3):345–370, 2002.

Jerome R Busemeyer and James T Townsend. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, 100(3): 432, 1993.

JR Busemeyer, S Gluth, J Rieskamp, and BM Turner. Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends Cogn Sci.*, 23(3):251–263, 2019.

J. Chen, R. Niedermeier, and P. Skowron. Stable marriage with multi-modal preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation (ACM:EC)*, pages 269–286. ACM, 2018.

F. Cooper and D. Manlove. Algorithms for new types of fair stable matchings. In *18th International Symposium on Experimental Algorithms, SEA 2020*, volume 160 of *LIPIcs*, pages 20:1–20:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020a.

Frances Cooper and David Manlove. Algorithms for new types of fair stable matchings. *arXiv preprint arXiv:2001.10875*, 2020b.

Joanna Drummond, Andrew Perrault, and Fahiem Bacchus. SAT is an effective and complete method for solving stable matching problems with couples. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, pages 518–525. AAAI Press, 2015.

I. Erev, E Ert, O Plonsky, D Cohen, and O Cohen. From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychol Rev.*, 124(4):369–409, 2017.

D. Gale and L. S. Shapley. College admissions and the stability of marriage. *Amer. Math. Monthly*, 69:9–14, 1962.

M Gelain, MS Pini, F Rossi, KB Venable, and T Walsh. Procedural fairness in stable marriage problems. In *10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, pages 1209–1210. IFAAMAS, 2011.

Mirco Gelain, Maria Silvia Pini, Francesca Rossi, Kristen Brent Venable, and Toby Walsh. Local search approaches in stable matching problems. *Algorithms*, 6(4):591–617, 2013.

Sushmita Gupta, Sanjukta Roy, Saket Saurabh, and Meirav Zehavi. Balanced stable marriage: How close is close enough? In *Workshop on Algorithms and Data Structures*, pages 423–437. Springer, 2019.

LLC Gurobi Optimization. Gurobi optimizer reference manual, 2020. URL http://www.gurobi.com.

Dan Gusfield and Robert W. Irving. *The Stable Marriage Problem: Structure and Algorithms*. MIT Press, Cambridge, MA, USA, 1989. ISBN 0262071185.

Pascal Van Hentenryck and Laurent Michel. *Constraint-based local search*. MIT Press, 2005.

Jared M Hotaling, Jerome R Busemeyer, and Jiyun Li. Theoretical developments in decision field theory: Comment on tsetsos, usher, and chater (2010). 2010.

Kazuo Iwama, Shuichi Miyazaki, and Hiroki Yanagisawa. Approximation algorithms for the sex-equal stable marriage problem. *ACM Trans. Algorithms*, 7(1):2:1–2:17, 2010.

F. Kojima and M.U. Unver. he "boston" school-choice mechanism: an axiomatic approach. *Econ Theory*, 55:515–544, 2014.

Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. Webuildai: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–35, 2019.

Seunghan Lee and Young Jun Son. Extending decision field theory to a multi-agent decision-making with forgetting. In *Proceedings of Institute of Industrial and Systems Engineers (IISE) Annual Conference 2016*, pages 2092–2097. IISE, 2016.

J. W. Lian, N. Mattei, R. Noble, and T. Walsh. The conference paper assignment problem: Using order weighted averages to assign indivisible goods. In *Proc. of the 33rd AAAI Conference*, 2018.

A. Loreggia, N. Mattei, F. Rossi, and K. B. Venable. Preferences and ethical principles in decision making. In *Proc. of the 1st AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2018.

David F. Manlove. *Algorithmics of Matching Under Preferences*, volume 2 of *Series on Theoretical Computer Science*. WorldScientific, 2013. ISBN 978-981-4425-24-7. doi: 10.1142/8591. URL https://doi.org/10.1142/8591.

Andrea Martin and Kristen Brent Venable. Decision making over combinatorially-structured domains. In *11th Multidisciplinary Workshop on Preferences Handling (MPREF 2018) co-located with AAAI 2018, New Orleans, LA, USA*, 2018.

Nicholas Mattei. Closing the loop: Bringing humans into empirical computational social choice and preference reasoning. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 5169–5173. ijcai.org, 2020. doi: 10.24963/ijcai.2020/729. URL https://doi.org/10.24963/ijcai.2020/729.

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

S. Miyazaki and K. Okamoto. Jointly stable matchings. *J Comb Optim*, 38:646–665, 2019.

William Pettersson, Maxence Delorme, Sergio García, Jacek Gondzio, Joerg Kalcsics, and David Manlove. Improving solve times of stable matching problems through preprocessing. *Computers and Operations Research*, 128:105–128, 2021.

R. Roe, J.R. Busemeyer, and J.T. Townsend. Multi-alternative decision field theory: A dynamic connectionist model of decision-making. *Psychological Review*, 108:370–392, 2001a.

Robert M Roe, Jermone R Busemeyer, and James T Townsend. Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological review*, 108 (2):370, 2001b.

F. Rossi and N. Mattei. Building ethically bounded AI. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI*, pages 9785–9789. AAAI Press, 2019.

Alvin E Roth. *Who Gets What – and Why: The New Economics of Matchmaking and Market Design.* Houghton Mifflin Harcourt, 2015.

Alvin E Roth, Uriel G Rothblum, and John H Vande Vate. Stable matchings, optimal assignments, and linear programming. *Mathematics of operations research*, 18(4):803–828, 1993.

T. Sühr, A.J. Biega, M. Zehlike, K.P. Gummadi, and A. Chakraborty. 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*, pages 3082–3092. ACM, 2019.

Jennifer S Trueblood, Scott D Brown, and Andrew Heathcote. The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological review*, 121(2):179, 2014.

N. Tziavelis, I. Giannakopoulos, R. Quist Johansen, K. Doka, N. Koziris, and P. Karras. Fair procedures for fair stable marriage outcomes. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020,*, pages 7269–7276. AAAI Press, 2020.

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 457–467, 2019.