

# THE REVEALED PREFERENCE THEORY OF STABLE AND EXTREMAL STABLE MATCHINGS

FEDERICO ECHENIQUE, SANGMOK LEE, MATTHEW SHUM, AND M. BUMIN YENMEZ

ABSTRACT. We investigate the testable implications of the theory of stable matchings. We provide a characterization of the data that are rationalizable as a stable matching when agents' preferences are unobserved. The characterization is a simple nonparametric test for stability, in the tradition of revealed preference tests. We also characterize the observed stable matchings when monetary transfers are allowed, and the stable matchings that are best for one side of the market (extremal stable matchings). We find that the theory of extremal stable matchings is observationally equivalent to requiring that there be a unique stable matching, or that the matching be consistent with unrestricted monetary transfers. We also present results on rationalizing a matching as the median stable matching.

## 1. INTRODUCTION

This paper studies the testable implications of *stability* in two-sided matching markets. Specifically, if one can observe matchings, but not agents' preferences, what are the observable implications of the theory? The paper develops such a theory of revealed preference for matching markets.

The revealed preference problem in matching presents unique challenges. In classical revealed-preference theory, if Catherine chooses option A over option B, then we may infer that she prefers A over B. In a two-sided model the situation is much more complicated: If Catherine matches with Jules and not with Jim, it may be because she likes Jules best, but it may also be because Jim is matched to someone *he* prefers over Catherine. Jim's preferences, however, are as unobservable as Catherine's. Hence, matching data do not unambiguously resolve the direction of revealed preference. This problem is a crucial challenge; it is intrinsic to two-sided models; and most empirical studies of matching circumvent the problem by assuming unlimited transfers among the agents.<sup>1</sup> Ours is the first paper to provide a complete revealed preference characterization in the absence of transfers.

---

This paper evolved from Echenique, Lee, and Yenmez (2010) and contains generalizations of the theoretical results in Echenique, Lee, and Shum (2010) both of which are obsolete now. We thank Lars Ehlers for questions that motivated some of the current research. We also thank participants and discussants at many universities and conferences. Echenique, Lee, and Shum are affiliated with the Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125; Yenmez is with Microsoft Research New England, One Memorial Drive, Cambridge, MA 02142 and Tepper School of Business, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213. Emails: [fede@caltech.edu](mailto:fede@caltech.edu), [sangmok@hss.caltech.edu](mailto:sangmok@hss.caltech.edu), [mshum@caltech.edu](mailto:mshum@caltech.edu), [byenmez@andrew.cmu.edu](mailto:byenmez@andrew.cmu.edu).

<sup>1</sup>Under this assumption, they can focus on the matchings that maximize social surplus. See Section 4.2.

The literature on stable matching has grown rapidly, including many theoretical advancements and refinements. It has been very successful as a normatively applicable theory, and guided the design of important matching markets (Roth, 2008; Sönmez and Ünver, Forthcoming). However, stable matching has yet to be well understood as a positive theory.

For example, much is known about the two canonical models of matching: the model with no monetary transfers (the marriage, or college-admissions model), and the model with transfers (the assignment game); but we do not know how to *empirically* distinguish one model from the other. To say that a matching conforms to one of these models requires information on agents' preferences, and these are unobservable. It is important to know if stability explains observed data on matching, and if monetary transfers play a role.

Another example relates to extremal stable matchings. In the absence of transfers, there are two *extremal* stable matchings: each one is the best for one side of the market, and the worst for the other side, among all stable matchings. These are theoretically appealing and practically useful. Centralized market clearinghouses, such as the large National Resident Matching Program in the U.S. (or the recent designs of matching of students to public schools), implement an extremal matching. This begs the question of whether the market acting in a decentralized fashion would also implement an extremal matching.<sup>2</sup> Since preferences are unobserved, we do not know when decentralized matchings are consistent with an extremal stable matching. Our paper presents such tests, nonparametric tests in the revealed preference tradition. Ours are tests for stability with transfers and without transfers, and for extremal stability.

Any study of the testable implications of a theory comes with an assumption about the kinds of data one may be able to observe. We focus on a general notion of data, which includes “aggregate matchings” as its main special case. In an aggregate matching, individuals on each side of the market are summed up into cells on the basis of their observed characteristics, such as age, educational attainment, or employment sector. This kind of data is entirely realistic: e.g. empirical researchers studying marriage matching typically use aggregated data (Choo and Siow, 2006). In some cases, disaggregated data is simply unavailable, for example in the application to Add-Health data we outline in Section 7. With aggregate matching data, we assume that all individuals with the same characteristics are identical, and have identical preferences. This is a strong assumption, but without it the theory has no testable implications: that is, any matching could be trivially rationalizable, once one allows for enough heterogeneity in preferences among observationally identical individuals. To eliminate these possibilities, we focus on the restrictive case of no preference heterogeneity

---

<sup>2</sup>Experimental evidence suggests that they do not. The decentralized experimental markets in Echenique and Yariv (2010) result in a stable matching, but it is usually a compromise between the two extremes (the median stable matching; see Section 6).

among individuals with the same characteristics. This can be considered a “worst case” under which rationalizability is still possible.<sup>3</sup>

Our notion of data is more general than just aggregate matchings. Given a randomized matching of agents to objects (a probabilistic assignment of children to schools, for example, as in Abdulkadiroğlu, Pathak, Roth, and Sönmez (2005) and Abdulkadiroğlu, Pathak, and Roth (2005)), one may ask if the randomized matching is consistent with stability for some preferences of the children and priorities of the schools. Our results are applicable to this randomized matching environment as well.

To lay some necessary groundwork, we first study stability in a general model of population matching; including, as special cases, aggregate (Choo and Siow, 2006; Dagsvik, 2000), and random matching (Hylland and Zeckhauser, 1979; Roth, Rothblum, and Vande Vate, 1993; Alkan and Gale, 2003). In doing so, we extend the basic theory of stable matching to our general model; including proofs of the existence and polarity properties of stable matchings. These results are important because, in order to study stable matchings, we must first establish that they exist and behave in the usual ways.<sup>4</sup> In particular, we show the existence of extremal and median stable matchings.

**Our main results are as follows:** First we characterize matchings that are *rationalizable*. A matching is rationalizable if there exist preferences for which the matching is stable. We represent a matching as a matrix  $X$ , where entry  $x_{i,j}$  can be interpreted as the number of agents of type  $i$  who match with agents of type  $j$ , or the probability that agent  $i$  matches with agent  $j$ . Then one can draw a graph on the matrix by having an edge (link) between two entries if they are not zero and lie on the same column or row of the matrix. Our main result is that a matching is rationalizable if and only if this graph has no two connected cycles.

Turning to matching with transfers: We can rationalize a matching as stable with transfers if and only if the above-mentioned graph has no cycles. So the model with transfers is empirically strictly stronger than the model without transfers. The models are nested, and it is possible to test for the existence of transfers under the assumption of stability.

Note that the tests have a very simple form and can some times (as in Section 7) be carried out with paper and pencil. Given a matching table, draw a graph by joining non-zero entries on the same row or column, and look for cycles.

---

<sup>3</sup>Moreover, such a strong assumption also needs to be relaxed when takes our approach to data. A similar situation arises in classical revealed preference theory, where the theory assumes that all observations are generated under stable preferences. When the theory is applied to actual data, however, one needs to introduce an error structure judiciously in order to allow for realistic shocks in the empirical modelling while, at the same time, not rendering the revealed preference problem trivial. See, for example, Varian (1985), Blundell, Browning, and Crawford (2003) or Echenique, Lee, and Shum (2011).

<sup>4</sup>Strictly speaking we work with the notion of strongly stable matching (see our discussion in Section 8.2), and so we need to extend the classical results on stability to strongly stable matchings.

Next we characterize matchings that are rationalizable as an extremal stable matching. We show that a matching is extremal-rationalizable if and only if the graph has no cycles, the same condition for rationalizability with transfers. It turns out that rationalizability as the unique stable matching also requires the same condition. Hence, the empirical content of the hypotheses that a marriage matching is man-optimal or woman-optimal is the same as the hypothesis that the matching is uniquely stable, or that there are unlimited quasilinear monetary transfers in the market.

In other words, the theory of extremal stable matching, which is more restrictive than the theory of stable matching, is observationally equivalent to the theory of unique stable matching; and observationally equivalent to the theory of transferable-utility matching. To gain some perspective on this finding, recall that the theory of stable matching without transfers was first developed by Gale and Shapley (1962), and generalized by Kelso and Crawford (1982); the algorithm developed by Gale and Shapley selects an extremal stable matching. On the other hand, the theory of stable matching with transfers was developed by Shapley and Shubik (1971), and (rather famously) applied to marriage matchings by Becker (1973). There are no reasons to *a priori* expect one theory to be more relevant than the other. *Our results imply that from a purely empirical viewpoint, the matching theory with transfers is nested in the matching theory without transfers; and the predictions of the Gale-Shapley algorithm are equivalent to Becker's model of marriage with transfers.*

Finally, we provide sufficient, but not necessary, conditions for median rationalizability. The difficulty of characterizing matchings that are median-rationalizable is that the whole set of stable matchings must be found to establish the median property.

We present a simple empirical illustration in Section 7. For the illustration, we use data on teenage dating which comes naturally in an aggregate form.

**1.1. Related literature.** A number of papers study the empirical content of stability for aggregate matching (Choo and Siow, 2006; Dagsvik, 2000). Ours is the first revealed preference results for aggregate matchings without transfers. We are also the first to embed empirically the matching model with transfers in the model without transfers.

Random and fractional matchings are studied by Vande Vate (1989); Rothblum (1992); Roth, Rothblum, and Vande Vate (1993); Kesten and Ünver (2009), but not from the revealed preference perspective. With the exception of Kesten and Ünver (2009), these papers focus on the standard notion of stability for fractional matching. Roth, Rothblum, and Vande Vate (1993) introduce the idea of strong stability, as we use it, but obtain only very weak results. The problem is that strong stability gives rise to a system of quadratic equations rather than a linear programming problem used in Vande Vate (1989); Rothblum (1992); Roth, Rothblum, and Vande Vate (1993) (see Section 8.2). Our existence results are closer to Alkan and Gale (2003), who present a general approach to stable matching. Alkan and

Gale work with choice functions instead of the incomplete first-order stochastic dominance preference relation that we use, but it is likely that their methods can also be adapted to show existence of extremal stable matchings in our model. Of course, Alkan and Gale do not treat the issue of rationalizing a matching when preferences are unknown.

The econometric problem of identifying and estimating preferences in the model with transfers is treated in, among others, Choo and Siow (2006), Dagsvik (2000), Fox (2007), and Galichon and Salanie (2009). Echenique (2008) and Chambers and Echenique (2009) deal with the revealed preference problem for a collection of individual-level matchings. They assume that there are many observations of matchings among the same agents; so they work with a different kind of matching data. Their results do not apply to aggregate or random matchings, and the results in these papers are totally unrelated to our results.

We use Tarski's fixed point theorem to show the existence of stable matchings and extremal stable matchings. This approach has been used in the matching literature before, for example Roth and Sotomayor (1988); Adachi (2000); Fleiner (2003); Echenique and Oviedo (2004, 2006); Echenique and Yenmez (2007); Ostrovsky (2008); Hatfield and Milgrom (2005); Küçükşenel (2011); Komornik, Komornik, and Viauroux (2010).

## 2. PRELIMINARY DEFINITIONS

**2.1. Lattice theory.** If  $S$  is a set and  $\leq$  is a partial order on  $S$ , we say that the pair  $(S, \leq)$  is a *partially ordered set*. We say that  $x, y \in S$  are *comparable* if  $x \leq y$  or  $y \leq x$ . A partially ordered set  $(S, \leq)$  is a *lattice* if, for every  $x, y \in S$ , the least upper bound and the greatest lower bound of  $\{x, y\}$  with respect to the partial order  $\leq$  exist in  $S$ . We denote the least upper bound of  $\{x, y\}$  by  $x \vee y$ ; and the greatest lower bound of  $\{x, y\}$  by  $x \wedge y$ . Similarly, if for every  $S' \subseteq S$ , the least upper bound and the greatest lower bound of  $S'$  exist in  $S$ , the lattice  $(S, \leq)$  is called *complete*. We say a lattice  $(S, \leq)$  is *distributive* if the following holds for all  $x, y, z \in S$ :

- $x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$ , and
- $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$ .

**2.2. Graph theory.** An (undirected) *graph* is a pair  $G = (V, L)$ , where  $V$  is a set and  $L$  is a subset of  $V \times V$ . A *path* in  $G$  is a sequence  $p = \langle v_0, \dots, v_N \rangle$  such that for  $n \in \{0, \dots, N-1\}$ ,  $(v_n, v_{n+1}) \in L$ . We write  $v \in p$  to denote that  $v$  is a vertex in  $p$ . A path  $\langle v_0, \dots, v_N \rangle$  *connects* the vertices  $v_0$  and  $v_N$ . A path  $\langle v_0, \dots, v_N \rangle$  is *minimal* if there is no proper subsequence of  $\langle v_0, \dots, v_N \rangle$  which is also a path connecting the vertices  $v_0$  and  $v_N$ . The *length* of path  $\langle v_0, \dots, v_N \rangle$  is  $N$ .

A *cycle* in  $G$  is a path  $c = \langle v_0, \dots, v_N \rangle$  with  $v_0 = v_N$ . A cycle is *minimal* if for any two vertices  $v_n$  and  $v_{n'}$  in  $c$ , the paths in  $c$  from  $v_n$  to  $v_{n'}$  and from  $v_{n'}$  to  $v_n$  are distinct and minimal. We call  $v$  and  $w$  *adjacent in*  $c$  if there is  $n$  such that  $v_n = v$  and  $v_{n+1} = w$  or

$v_n = w$  and  $v_{n+1} = v$ . If  $c$  and  $c'$  are two cycles, and there is a path from a vertex of  $c$  to a vertex of  $c'$ , then we say that  $c$  and  $c'$  are **connected**.

### 3. MODEL

**3.1. Matching with non-transferable utilities (NTU).** The primitives of the model are represented by a four-tuple  $\langle M, W, P, K \rangle$ , where

- $M$  and  $W$  are finite and disjoint sets of, respectively **types of men**, and **types of women**.
- $P$  is a **preference profile**: a list of preferences  $P_m$  for every type of man  $m$  and  $P_w$  for every type of woman  $w$ . Each  $P_m$  is a linear order over  $W \cup \{w_0\}$ , and each  $P_w$  is a linear order over  $M \cup \{m_0\}$ . Here,  $w_0$  and  $m_0$  represent the alternative of being unmatched. The weak order associated with  $P_m$  ( $P_w$ ) is denoted by  $R_m$  ( $R_w$ ).
- $K$  is a list of non-negative real numbers  $K_m$  for each  $m \in M$  and  $K_w$  for each  $w \in W$ . There are  $K_m$  men of type  $m$  and  $K_w$  women of type  $w$ .

We enumerate  $M$  as  $\{m_1, \dots, m_{|M|}\}$  and  $W$  as  $\{w_1, \dots, w_{|W|}\}$ . A **matching** is a  $|M| \times |W|$  matrix  $X = (x_{m,w})$  such that  $x_{m,w} \in \mathbf{R}_+$ ,  $\sum_w x_{m,w} \leq K_m$  for all  $m$ , and  $\sum_m x_{m,w} \leq K_w$  for all  $w$ .<sup>5</sup> We denote the mass of single  $m$ -agents in  $X$  by  $x_{m,0}$ , and denote the mass of single  $w$ -agents in  $X$  by  $x_{0,w}$ .

$X_{m_i, \cdot}$  is the  $i^{\text{th}}$  row and  $X_{\cdot, w_j}$  is the  $j^{\text{th}}$  column. When it is not ambiguous, we write  $X_{m_i}$  or  $X_i$  for  $X_{m_i, \cdot}$ , and  $X_{w_j}$  or  $X_j$  for  $X_{\cdot, w_j}$ . Similarly, we use  $x_{i,j}$  for  $x_{m_i, w_j}$ .

**Definition 1.** A matching  $X$  is **individually rational** if  $x_{m,w} > 0$  implies that  $w R_m w_0$  and  $m R_w m_0$ .

A pair  $(m, w)$  is a **blocking pair** for  $X$  if there is  $m'$  and  $w'$  such that  $m P_w m'$ ,  $w P_m w'$ ,  $x_{m,w'} > 0$ , and  $x_{m',w} > 0$ .

A matching  $X$  is **stable** if it is individually rational and there are no blocking pairs for  $X$ .

We denote by  $S(M, W, P, K)$  the set of all stable matchings in  $\langle M, W, P, K \rangle$ .

Two special cases of our model are worth emphasizing: The model of **random matching** is obtained when  $K_m = 1$  for all  $m$ , and each  $K_w$  is a natural number. The interpretation of random matching is that men are “students” and women are “schools.” Students are assigned to schools at random, and each school  $w$  has  $K_w$  seats available for students. In real-life school choice, the randomization often results from indifferences in schools’ preferences over students (Abdulkadiroğlu, Pathak, and Roth, 2005): Matching theory often requires strict preferences, so a random “priority order” is produced for the schools in order to break indifferences. Random matchings arise in many other situations as well, because random

<sup>5</sup> $\mathbf{R}_+$  denotes the set of non-negative real numbers.

assignment is often a basic consequence of fairness considerations (if two children want the same toy, they may agree to flip a coin). Here we are mainly interested in situations where a random assignment is given in unambiguous terms, but preferences are unobserved. It is also possible that they are observed, but we suspect that they have been misrepresented or observed with errors.

The second model is that of *aggregate matching*, where all  $K_m$  and  $K_w$  are natural numbers, and all entries of a matching  $X$  are natural numbers. The interpretation is that there are  $K_m$  men of type  $m$ , and  $K_w$  women of type  $w$ , and that a matching  $X$  exhibits in  $x_{m,w}$  how many type  $m$  men matched to type  $w$  women. The model of aggregate matching captures actual observations in marriage models. We observe that men and women are partitioned into types according to their observable characteristics (age, income, education, etc.); and we are given a table showing how many type  $m$  men married type  $w$  women. These observations are essentially “flow” observations (marriages in a given year), so the aggregate matchings do not have any single agents.

An implication of these models is that we assume preferences to be at the type-specific level, rather than the individual-level. As we remarked earlier, without restrictions on preferences, we would lose all empirical consequences in our nontransferable utility setting. This differs from the approach in empirical applications of matching theory, which assume transferable utilities (see, e.g., Choo and Siow (2006) or Galichon and Salanie (2009)), but allow for heterogeneous preferences at the individual-level. When our results are taken to data, one needs to introduce an appropriate error structure that allows for individual heterogeneity. One way of doing this is by introducing measurement error, just as in the classical revealed preference theory for consumption (Varian, 1985). In Section 7, we present a simple illustration using data on high-school romantic relationships, where we take such an approach.

**3.2. Matching with transferable utilities (TU).** The primitives of the model are represented by a four-tuple  $\langle M, W, \mathcal{A}, K \rangle$ .  $M$ ,  $W$ , and  $K$  are defined the same as NTU model.

- $M$  and  $W$  are finite and disjoint sets of, respectively *types of men*, and *types of women*.
- $\mathcal{A} = (\alpha_{m,w})$  is a  $|M| \times |W|$  matrix of non-negative real numbers.  $\mathcal{A}$  is called a *surplus matrix*, in which  $\alpha_{m,w}$  is a surplus that a type  $m$  man and a type  $w$  woman jointly generate.
- $K$  is a list of non-negative real numbers  $K_m$  for each  $m \in M$  and  $K_w$  for each  $w \in W$ . There are  $K_m$  men of type  $m$  and  $K_w$  women of type  $w$ .

A *matching* is a  $|M| \times |W|$  matrix  $X = (x_{m,w})$  such that  $x_{m,w} \in \mathbf{R}_+$ ,  $\sum_w x_{m,w} \leq K_m$  for all  $m$ , and  $\sum_m x_{m,w} \leq K_w$  for all  $w$ . A *payoff* is a pair of  $|M| \times |W|$  matrices  $(U, V)$

such that  $(u, v)_{m,w} \in \mathbf{R}_+^2$  for all  $(m, w) \in M \times W$ .<sup>6</sup> A matching  $X$  generates a payoff  $(U, V)$  by  $u_{m,w} + v_{m,w} = \alpha_{m,w}x_{m,w}$ .

**Definition 2.** A matching  $X$ , generating a payoff  $(U, V)$ , is **stable** if

- $u_{m,w} \geq 0, v_{m,w} \geq 0$ ,
- $\frac{u_{m,w'}}{x_{m,w'}} + \frac{v_{m',w}}{x_{m',w}} \geq \alpha_{m,w}$  for all  $(m, w) \in M \times W$ ,  $m' \in M$ , and  $w' \in W$  such that  $x_{m,w'}x_{m',w} > 0$ .

The first condition reflects that an agent always has an option of remaining single. The second condition requires that the outcome is not blocked by any pair of type  $m$  men and type  $w$  women. When the condition is violated, the pair of type  $m$  men and type  $w$  women may form a blocking pair with breaking some of their partnerships with  $w'$  and  $m'$ .

On the other hand, consider the following problem.

$$(1) \quad \max_{X \in \mathbf{R}_+^{|M| \times |W|}} \sum_{m,w} \alpha_{m,w}x_{m,w} \quad s.t. \quad \begin{cases} \forall m \sum_w x_{m,w} \leq K_m \\ \forall w \sum_m x_{m,w} \leq K_w \end{cases}.$$

A matching  $X$  is called **optimal** if it is a solution of (1). An optimal matching achieves the most mutual surplus under a set of population constraints:  $K_m$  and  $K_w$ . It is well-known that the set of stable matchings is equivalent to the set of optimal matchings (Shapley and Shubik, 1971).

**Proposition 1.** A matching  $X$  is optimal if and only if it is stable.

The model of **assignment game** is obtained when  $K_m = 1$  for all  $m$ , and  $K_w = 1$  for all  $w$ . An interpretation of the assignment game is that men are “sellers” and women are “buyers.” When a bid is accepted, the buyer receives the object by paying money to the seller.

**3.3. Extremal stable matchings in the NTU model.** We now turn to extremal stable matchings in the NTU model. First, we establish that they exist for our general model of matchings. In fact, we show that the standard theory of the structure of stable matchings extends.<sup>7</sup>

<sup>6</sup>Note that agents of the same type may obtain different payoffs depending on the types of whom they matched with.

<sup>7</sup>As we explain in Section 8.2, since we work with strong stability, the existing results on stable – rather than strong stable – random matchings do not apply to our case. For the model of aggregate matching, there are no previous results on the structure of stable matchings.

For each  $m \in M$ , define

$$\mathcal{X}_m = \left\{ x \in \mathbf{R}_+^{|W|} : \sum_{1 \leq j \leq |W|} x_j \leq K_m \right\},$$

and a partial order  $\leq_m$  on  $\mathcal{X}_m$  as

$$y \leq_m x \quad \text{iff} \quad \forall w \in W \cup \{w_0\}, \quad \sum_{\substack{0 \leq j \leq |W| \\ w_j R_m w}} y_j \leq \sum_{\substack{0 \leq j \leq |W| \\ w_j R_m w}} x_j,$$

where  $x_0 = K_m - \sum_{1 \leq j \leq |W|} x_j$ , and  $y_0 = K_w - \sum_{1 \leq j \leq |W|} y_j$ .

Note that  $\leq_m$  is defined by analogy to the first order stochastic dominance order of probability distributions. In the case where  $K_m = 1$ , vectors  $x$  and  $y$  in  $\mathcal{X}_m$  represent probability distributions over the types of women that  $m$  may match to. In that case  $y \leq_m x$  if and only if the lottery induced by  $y$  over  $W \cup \{w_0\}$  is worse than the lottery induced by  $x$ , for any von Neumann-Morgenstern utility function. Letting  $\mathcal{X}_w = \left\{ x \in \mathbf{R}_+^{|M|} : \sum_{1 \leq i \leq |M|} x_i \leq K_w \right\}$ , we define  $\leq_w$  in an analogous way.

We introduce two partial orders on matchings. Suppose  $X$  and  $Y$  are matchings, then:

- $X \leq_M Y$  if, for all  $m$ ,  $X_m \leq_m Y_m$
- $X \leq_W Y$  if, for all  $w$ ,  $X_w \leq_w Y_w$ .

**Theorem 1.**  $(S(M, W, P, K), \leq_M)$  and  $(S(M, W, P, K), \leq_W)$  are nonempty, complete, and distributive lattices; in addition, for  $X, Y \in S(M, W, P, K)$

- (1)  $X \leq_M Y$  if and only if  $Y \leq_W X$ ;
- (2) for all  $a \in M \cup W$ , either  $X_a \leq_a Y_a$  or  $Y_a \leq_a X_a$ ;
- (3) for all  $m$ ,  $\sum_{w \in W} x_{m,w} = \sum_{w \in W} y_{m,w}$  and for all  $w$ ,  $\sum_{m \in M} x_{m,w} = \sum_{m \in M} y_{m,w}$ .

Theorem 1 presents versions of other classical results on matching for our model that we prove in Appendix A. Statement (1) is a ‘‘polarity of interests’’ results, saying that a stable matching  $X$  is better for men if and only if it is worse for women. Statement (2) says that the outcomes in two stable matchings are always comparable for an agent: note that  $\leq_a$  is an incomplete preference relation. Statement (3) is the ‘‘rural hospitals theorem,’’ which says that single agents in any two stable matchings are the same.

Theorem 1 also implies that there are two stable matchings,  $X^M$  and  $X^W$ , such that for all stable matchings  $X$ ,

$$\begin{aligned} X^W &\leq_M X \leq_M X^M, \text{ and} \\ X^M &\leq_W X \leq_W X^W. \end{aligned}$$

We refer to  $X^M$  as the **man-optimal** (M-optimal) stable matching, and to  $X^W$  as the **woman-optimal** (W-optimal) stable matching. We also call  $X^M$  and  $X^W$  **extremal** stable

matchings. A matching  $X$  is the unique stable matching if  $S(M, W, P, K) = \{X\}$ ; in this case  $X$  coincides with the  $M$ - and the  $W$ -optimal stable matchings.

#### 4. STATEMENT OF THE PROBLEM

We proceed to give formal statements of the problems we shall address. We give definitions and motivations for rationalizability without transfers, rationalizability with transfers, and extremal rationalizability. We assume that there are no single men or women. In practice, we observe only formed couples, and adding singles is qualitatively intact in our analysis.<sup>8</sup>

**4.1. Rationalizability.** Suppose that we are given a matching  $X$ , and the corresponding  $M$ ,  $W$  and  $K$ . We want to understand when there are preferences for the different types of men and women such that  $X$  is a stable matching. We say that a matching  $X$  is **rationalizable** if there exists a preference profile  $P = ((P_m)_{m \in M}, (P_w)_{w \in W})$  such that  $X$  is a stable matching in  $\langle M, W, P, K \rangle$ .

We present two simple observations to motivate our approach. The observations give a first view into the rationalizable matchings. Proposition 2 is subsumed in Theorem 2, but it is valuable as a motivation.

*Remark 1.* If  $|M| = |W| = 2$  then any matching  $X$  is rationalizable with the following preferences.

$$\begin{array}{cc|cc} P_{m_1} & P_{m_2} & P_{w_1} & P_{w_2} \\ \hline w_1 & w_2 & m_2 & m_1 \\ w_2 & w_1 & m_1 & m_2 \end{array}$$

**Proposition 2.** *If  $X$  has a  $2 \times 3$  or a  $3 \times 2$  submatrix that have positive elements, then  $X$  is not rationalizable.*

*Proof.* We may assume that  $X$  is the  $2 \times 3$  submatrix in question. Suppose  $X$  is stable for a preference profile  $P$ . By individual rationality, for all types of men any type of woman is preferable to being single and similarly for the types of women. There must be at least one pair  $(m, w)$  such that  $w$  is not last in  $m$ 's preference, and  $m$  is not last in  $w$ 's preferences. Finding this pair suffices because then there is  $m'$  and  $w'$  with  $x_{m,w'} > 0$  and  $x_{m',w} > 0$  and  $w P_m w'$ ,  $m P_w m'$ :  $(m, w)$  is a blocking pair. Say that  $m_1$  ranks  $w_1$  last. If either  $w_2$  or  $w_3$  rank  $m_1$  as not-last, then we are done. If both  $w_2$  and  $w_3$  rank  $m_1$  last then consider  $m_2$ :  $m_2$  must rank one of  $w_2$  and  $w_3$  as not-last. Since both  $w_2$  and  $w_3$  rank  $m_2$  as not-last, then we are done.  $\square$

<sup>8</sup>Add a column  $w_0$  and a row  $m_0$  to  $X$ . Let  $x_{m,w_0}$  be the number of type  $m$  men who are single,  $x_{m_0,w}$  the number of type  $w$  women who are single, and  $x_{m_0,w_0} = 0$ . A result similar to Theorem 2 holds for this augmented matrix.

Remark 1 demonstrates that a  $2 \times 2$  matrix of positive elements is rationalizable with a cycled preference profile: for instance, type  $m_1$  men prefer type  $w_1$  women, who prefer not type  $m_1$  but type  $m_2$  men, who in turn prefer not type  $w_2$  but type  $w_1$  women, etc. Whereas, Proposition 2 shows that such a way of constructing a preference profile fails once the matrix becomes larger. That is, rationalizability requires for a matrix to have relatively *sparse* positive elements: it cannot have too many non-zero elements.

**4.2. TU-rationalizability.** Suppose that we are given a matching  $X$ . We want to understand when there exists a surplus matrix  $\mathcal{A}$  and payoff matrices  $(U, V)$  such that  $X$  is stable. Considering Proposition 1, the set of stable matchings is observationally equivalent to the set of optimal matchings. Thus, with regards to rationalizability, we focus on the set of optimal matchings.

Formally, a matching  $X$  is **TU-rationalizable** by a matrix of surplus  $\mathcal{A}$  if  $X$  is the *unique* solution to the following problem:

$$(2) \quad \max_{\tilde{X} \in \mathbf{R}_+^{|M| \times |W|}} \sum_{m,w} \alpha_{m,w} \tilde{x}_{m,w}$$

$$s.t. \quad \begin{cases} \forall m \quad \sum_w \tilde{x}_{m,w} = \sum_w x_{m,w} = K_w \\ \forall w \quad \sum_m \tilde{x}_{m,w} = \sum_m x_{m,w} = K_m \end{cases} .$$

*Remark 2.* Essentially, we restrict attention to situations where the number of agents of each type is given, and we focus on how they match. The restriction is obviously needed, as one could otherwise generate high surplus by re-classifying agents into high-surplus types.

Note also that we require  $X$  to be the unique maximizer in (2). This contrasts with the definition of rationalizability without transfers, where we did not require  $X$  to be the unique stable matching. This difference is inevitable, though. If we instead required  $X$  to be only one of the maximizers of (2), then any matching could be rationalized with a constant surplus ( $\alpha_{m,w} = c$  for all  $m, w$ ). In a sense, without transfers multiplicity is unavoidable (only very strong conditions ensure a unique stable matching), while uniqueness in the TU model holds for almost all real matrices  $\mathcal{A}$ .

**4.3. Extremal-rationalizability.** Finally, we study the revealed preference question for extremal stable matchings. Formally, a matching  $X$  is **M-optimal rationalizable** if there is a preference profile  $P = ((P_m)_{m \in M}, (P_w)_{w \in W})$  such that  $X$  is the M-optimal stable matching in  $\langle M, W, P, K \rangle$ . Analogously, it is **W-optimal rationalizable** if there is a preference profile such that  $X$  is the W-optimal matching in the corresponding market.

## 5. MAIN RESULTS

We consider the graph formed by connecting any two non-zero elements of the matrix, as long as they lie on the same row or column. It turns out that the rationalizability of a matching depends on the structure of this graph. Formally, we associate to each matching  $X$  a graph  $(V, L)$  defined as follows. The set of vertices  $V$  is  $\{(m, w) : m \in M, w \in W \text{ such that } x_{m,w} > 0\}$ , and an edge  $((m, w), (m', w')) \in L$  is formed for every pair of vertices  $(m, w)$  and  $(m', w')$  with  $m = m'$  or  $w = w'$ .

The possibilities of edges in the matching graph – one type of man matched with more than one type of woman, and vice versa – is a crucial distinction between individual and aggregate matchings. To understand their importance for rationalizability, consider a one-to-one individual matching. Obviously, there are no edges here, and also the matching is trivially rationalizable: We can set preferences such that each agent’s match is his/her most preferred partner, so the observed matching will be stable for these preferences.

In an aggregate (or random) matching, however, an edge invalidates these preferences; indeed, for a “vertical” edge (in which two women of the same type marry men of different types), strict preferences imply that, at the least, some women of this type are *not* matched to their most preferred type of men. For this to be stable, it must be that more preferable types of men are “not available” to these women – that is, these men are matched to women whom these men find more preferable. Obviously, this imposes restrictions on preferences. In the presence of edges, then, the rationalizability question boils down, essentially, to the *number* and *configurations* of edges which can be allowed for, such that one can still devise preferences consistent with all the restrictions implied by the edges. Recall the intuition presented in the introduction, involving Catherine, Jules and Jim: by translating the question into the graph  $G$  we can get a handle on the problem of the circularity of revealed preferences. The rationalizability results presented below answer this question explicitly.

**Theorem 2** (Rationalizability). *A matching  $X$  is rationalizable if and only if the associated graph  $(V, L)$  does not contain two connected distinct minimal cycles.*

The following example illustrates the condition in Theorem 2.

*Example 1* (minimal cycle). Let  $X$  be

$$\begin{pmatrix} 11 & 9 & 10 \\ 0 & 22 & 41 \\ 13 & 91 & 0 \end{pmatrix}.$$

The graph  $(V, L)$  can be represented as

$$\begin{array}{c} 11 \text{ --- } 9 \text{ --- } 10 \\ \left( \begin{array}{c} 0 \\ \left( \begin{array}{c} | \\ 22 \text{ --- } 41 \\ | \end{array} \right) \\ 13 \text{ --- } 91 \end{array} \right) \quad 0 \end{array}$$

The following is an example of two minimal cycles that are connected:

$$\begin{array}{c} 11 \text{ --- } 9 \quad 10 \\ \left( \begin{array}{c} 0 \\ \left( \begin{array}{c} | \\ 22 \quad 41 \\ | \end{array} \right) \\ 13 \text{ --- } 91 \end{array} \right) \quad 0 \end{array} \quad \begin{array}{c} 11 \text{ --- } 9 \text{ --- } 10 \\ \left( \begin{array}{c} 0 \\ \left( \begin{array}{c} | \\ 22 \text{ --- } 41 \\ | \end{array} \right) \\ 13 \text{ --- } 91 \end{array} \right) \quad 0 \end{array}$$

While the full proof is in Appendix B, we provide some intuition here. As discussed above, edges in an observed matching impose restrictions on preferences. Cycles, then, which are formed of connected edges, impose even more restrictions. Consider, for example, the left-hand side cycle presented in Example 1. This cycle consists of four edges connecting two types of men (call them  $m_1$  and  $m_2$ ), and two types of women (call them  $w_1$  and  $w_2$ ). In this cycle, men of type  $m_1$  are matched to women of both types  $w_1$  and  $w_2$ . Because of strict preferences, however,  $m_1$  cannot be indifferent between these two types of women. Assuming that women of type  $w_1$  are more preferred (ie,  $w_1 P_{m_1} w_2$ ), then, it must be that for the men of type  $m_1$  who are matched to  $w_2$ , the preferable women of type  $w_1$  are not “available” to him – specifically, women of type  $w_1$  who are matched with men of type  $m_2$  must prefer their spouse to men of type  $m_1$  (that is,  $m_2 P_{w_1} m_1$ ). Obviously, repeating this argument for all four edges in the cycle leads to a very large number of restrictions on the preferences between the types of men and women in the cycle. It turns out (via Lemmas 4 and 5 in Appendix B) that there are only *two possible sets* of preferences among the four types in the cycle which are consistent with stability– cycles, then, do severely restrict preferences.

In this light, then, Theorem 2 says that multiple cycles can coexist in a stable matching only if they are *isolated* – that is, there is no path composed of edges which connect the cycles. Intuitively, this is because when cycles are isolated, the preferences among the types in one cycle do not affect the preferences in another cycle. However, when cycles are connected, then preferences among the types in these cycles are interdependent, and may not be mutually coherent. Indeed, Theorem 2 states (and the proof of the necessary part makes explicit) that even with *two* connected minimal cycles, the restrictions on preferences imposed by the first cycle are inconsistent with the restrictions on preferences imposed by the second cycle.

The proof of the sufficiency direction of Theorem 2 is constructive; it works by using an algorithm to construct a rationalizing preference profile. The construction is not universal, in the sense that some rationalizing preference profiles cannot be constructed using the

algorithm (see Example F.1; there are typically multiple rationalizing preferences, if there are any).

The next result deals with conditions for extremal rationalizability – whether preferences can be found such that an observed matching can be rationalized as a man or woman-optimal matching.

**Theorem 3** (Extremal-rationalizability). *Let  $X$  be a matching. The following statements are equivalent:*

- (1)  $X$  is  $M$ -optimal rationalizable;
- (2)  $X$  is  $W$ -optimal rationalizable;;
- (3)  $X$  is rationalizable as the unique stable matching;
- (4) the graph  $(V, L)$  associated to  $X$  has no minimal cycles.

The proof is in Appendix D.

*Example 1 (continued):* Comparing the two previous theorems, an extremal matching imposes additional empirical restrictions relative to non-extremal stable matchings. To understand these additional restrictions, consider again the matching from Example 1, and assume (contradicting Theorem 3) that the matching in that example is  $M$ -optimal. Focus on the left-hand side cycle from that example. There, the horizontal edge, along with the assumption of strict preferences, implies that  $m_1$  is not indifferent between  $w_1$  and  $w_2$ ; assume, then, that man  $m_1$  prefers  $w_2$  to  $w_1$ . Therefore, shifting one type  $m_1$  man from  $w_1$  to  $w_2$  would make this man better off. However, in order to be a matching, this shift must be accompanied by other changes elsewhere in the matching.  $M$ -optimal rationalizability of the original matching implies that there are no other changes which are possible, which would not lead to some other men being matched to less preferable women. However, when there is a cycle, then it turns out that such a change is in fact possible – indeed, by Lemmas 4 and 5 in Appendix B, if preferences satisfy  $w_2 P_{m_1} w_1$ , then preferences must also satisfy  $w_1 P_{m_2} w_2$ , so that shifting one type- $m_1$  man from  $w_1$  to  $w_2$  while, at the same time, shifting one type- $m_2$  man from  $w_2$  to  $w_1$  would increase the aggregate utility of the types  $m_1$  and  $m_2$  men, while leaving the utility of all other types of men undisturbed. This perturbed matching is

$$\begin{pmatrix} 10 & 10 & 10 \\ 0 & 22 & 41 \\ 14 & 90 & 0 \end{pmatrix}.$$

This overall improvement in the aggregate utility of the men obviously contradicts the assumed  $M$ -optimality of the original matching, and illustrates why  $M$ -optimal rationalizability rules out cycles.

The final result is on TU-rationalizability:

**Theorem 4** (TU-rationalizability). *A matching  $X$  is TU-rationalizable if and only if the associated graph  $(V, L)$  contains no minimal cycles.*

The proof is in Appendix C. Thus, the theory of stability in the presence of transfers is empirically strictly stronger than stability in the model without transfers.

**Corollary 1.** *If a matching  $X$  is TU-rationalizable, then it is rationalizable.*

Theorems 2-4 provide a complete picture of the empirical content of models of stable matching. There are simple empirical tests of stability. The tests imply that the matching model with transfers is nested in the model without transfers, and that a stable matching with transfers is observationally equivalent to extremal (and unique) stable matchings without transfers, which is stated as a corollary below.

**Corollary 2.** *A matching  $X$  is  $M$ -optimal (or  $W$ -optimal) rationalizable if and only if it is TU-rationalizable.*

## 6. MEDIAN STABLE MATCHING: EXISTENCE AND TESTABLE IMPLICATIONS

Perhaps extremal stable matchings are unreasonable because they favor one side of the market over the other. One may instead be interested in matchings that present a compromise. The median stable matching provides such a compromise by assigning each agent a partner who is median ranked amongst all of his/her stable matching partners. Indeed in an experimental study on decentralized market, Echenique and Yariv (2010) find that the median stable matching tends to emerge among stable matchings. As such, we may want to test whether an observed matching in field data is rationalizable as a median stable matching.

It is not obvious that median stable matchings exist: for the standard models of matching, existence was proven by Teo and Sethuraman (1998); see also Klaus and Klijn (2010); and Schwarz and Yenmez (2011) for other matching markets. We shall first show that median stable matchings exist by only considering aggregate matchings, and then present a sufficient condition for rationalizability as a median stable matching.

We consider a market  $\langle M, W, P, K \rangle$ , where all numbers  $K_m, K_w$ , and entries of matchings  $X$  are non-negative integers. Subsequently, the number of stable matchings is finite, say  $n$ . Let  $S(M, W, P, K) = \{X^1, \dots, X^n\}$  be the set of stable matchings. For each agent  $a$  we consider all stable outcomes and rank them according to  $\geq_a$ . All the outcomes are comparable by (2) of Theorem 1. More formally, let  $\{X_a^{(1)}, \dots, X_a^{(n)}\} = \{X_a^1, \dots, X_a^n\}$  and  $X_a^{(1)} \geq_a \dots \geq_a X_a^{(n)}$ . Using these ranked outcomes we construct the following matrices:  $Y_m^{(l)} = X_m^{(l)}$  and  $Y_w^{(l)} = X_w^{(n+1-l)}$  for all  $l = 1, \dots, n$ . Matrix  $Y^{(l)}$  assigns each type of men the  $l^{\text{th}}$  best outcome, and each type of women the  $l^{\text{th}}$  worst outcome among stable matching partners.

**Proposition 3.**  $Y^{(l)}$  is a stable matching.

The proof follows from lattice structure with the operators  $\vee$  and  $\wedge$ , and essentially the same as the proofs of Theorem 3.2 in Klaus and Klijn (2006) and Theorem 1 of Schwarz and Yenmez (2011).

If  $n$  is odd we term  $Y^{(\frac{n+1}{2})}$  the **median stable matching**. If  $n$  is even we refer to  $Y^{(\frac{n}{2})}$  (or  $Y^{(\frac{n}{2}+1)}$ ) as the median stable matching (of course the choice is arbitrary).

**Corollary 3.** *The median stable matching exists.*

Having established that the median stable matching always exists in the model of aggregate matching, we proceed to understanding their testable implications. We want to know when a matching can be rationalized as a median stable matching. Formally, an aggregate matching  $X$  is **median-rationalizable** if there is a preference profile  $P$  such that  $X$  is the median stable matching in  $\langle M, W, P, K \rangle$ . Median-rationalizability depends on the number  $x_{m,w}$  as well as whether  $x_{m,w}$  is positive or 0.

If  $\langle v_0, \dots, v_N \rangle$  is a minimal cycle, then  $N$  is an even number. We say that a minimal cycle  $c$  is **balanced** if

$$\min \{v_0, v_2, \dots, v_{N-2}\} = \min \{v_1, v_3, \dots, v_{N-1}\}.$$

**Theorem 5.** *An aggregate matching  $X$  is median-rationalizable if it is rationalizable and if all minimal cycles of the associated graph  $(V, L)$  are balanced.*

The proof is in Appendix E.<sup>9</sup>

Note that the stability and rationalizability of a matching only depend on which entries are zero or positive. Thus if an observed matching  $X$  is *canonicalized* into  $X^C$  ( $x_{m,w}^c = 0$  if  $x_{m,w} = 0$ , and  $x_{m,w}^c = 1$  if  $x_{m,w} > 0$ ), then obviously  $X$  is rationalizable if and only if  $X^C$  is rationalizable. Given the role of canonical matchings in rationalizability, the following corollary is of some interest.

**Corollary 4.** *A canonicalized matching  $X^C$  is either not rationalizable or it is median-rationalizable.*

Unfortunately, the result in Theorem 5 is only a sufficient condition for median-rationalizability. Corollary 4 gives a characterization of necessary conditions, but only for the case of canonical matchings. To sketch the boundaries of these results, we present two examples in Appendix F. Example F.2 shows that there are indeed matchings that are not rationalizable as median stable matchings, so Corollary 4 on canonical matchings does not extend to all aggregate matchings. Example F.3 shows that the sufficient condition in Theorem 5 is not necessary.

<sup>9</sup>In our proofs we construct preferences such that the resulting number of stable matchings is odd. Therefore, our results do not depend on the choice of  $Y^{(\frac{n}{2})}$  or  $Y^{(\frac{n}{2}+1)}$  when  $n$  is even.

## 7. EMPIRICAL ILLUSTRATION: HIGH SCHOOL ROMANTIC RELATIONSHIPS

To illustrate the empirical relevance of our results, we apply them to data from the National Longitudinal Adolescent Health Data (Add Health), which has been used extensively in studies of social networks (e.g. Bearman, Moody, and Stovel (2004); Currarini, Jackson, and Pin (2009)). We focus on “romantic relationships” among teenage students, and use our results to test for stability.

The Add Health data fits our framework well because only the aggregate (school-level) matching of romantic relationships is observed. Specifically, the data records detailed information about high school friendships and a romantic relationship network; however, for confidentiality reasons, the names and identification numbers for the romantic partners are not provided. Therefore, at the school-level, the matching table is only known up to the demographic variables of the partners; as such, we do not identify individual level relationships, but only observe an aggregate matching embodying the number of romantic couples with different demographic types.

We use data from the ‘Adolescent In-Home Questionnaire’, in which students are asked about their romantic relationships. Amongst 20,745 survey respondents, 3,910 students have relatively stable on-going relationships at the time of the survey.<sup>10</sup> Furthermore, we restrict our attention to couples involving white students.<sup>11</sup> In all, we constructed aggregate matchings for 39 schools in which there were 20 or more couples, and defined the types of boys and girls according to their age.

We proceed to show three selected matchings to illustrate how we use our results. First, consider the matching in Table 1. By connecting positive entries that lie on the same row or column, we construct a graph. The graph has *one* cycle: the aggregate matching is rationalizable, but not TU-rationalizable, nor extremal (or unique) rationalizable (see Theorems 2 and 3).

Next consider Table 2, which contains two matchings which are neither rationalizable nor TU-rationalizable – that is, they contain more than two connected cycles. Indeed, out of the 39 schools, most of them have aggregate matchings which are not rationalizable at all.

One way to proceed in this case is to sequentially eliminate “rare” couples. For example, in the matching matrix from school #33, if we eliminate all entries with just one couple (2.6% of total number of couples), then all cycles are eliminated, and the matching becomes both TU-rationalizable as well as rationalizable. For the matching from school #47, however, we need to eliminate all entries with six (5.8%) or fewer couples, in order to make the

<sup>10</sup>We disregard relationships beginning at the year of survey (1995) as possibly *unstable* relationships; therefore, the relationships we retain have lasted at least four months.

<sup>11</sup>Because the students in our sample are predominantly white, and there are very few interracial couples, adding the non-white and interracial couples would not change the results, but only complicate the analysis by adding additional rows and columns to the matching table for the different racial groups.

Age $\sigma \downarrow, \varphi \rightarrow$	School ID: 19				
	-15	16	17	18	19-
-15	0	1	0	0	0
16	1	5	0	0	0
17	2	2	0	0	0
18	0	0	2	0	0
19-	0	0	4	5	3

TABLE 1. A rationalizable, but not TU/Extremal/Unique-rationalizable matching.

matching rationalizable. In this way, the matching in school #33 appears closer to being rationalizable than the matching in school #47. One interpretation of this procedure is that the observations are subject to measurement error: we may regard small entries in a matrix as “noisy” misclassified entries; and omit these for the purposes of testing for stability.<sup>12</sup>

Age $\sigma \downarrow, \varphi \rightarrow$	School ID: 33					School ID: 47				
	-15	16	17	18	19-	-15	16	17	18	19-
-15	(1)	0	0	0	0	0	(2)	0	0	(1)
16	0	0	(1)	0	0	(2)	7	(3)	0	0
17	4	3	0	3	0	(3)	(3)	(6)	(5)	0
18	(1)	(1)	6	6	4	(4)	(5)	8	9	7
19-	0	0	(1)	6	(1)	(2)	7	8	14	(6)

TABLE 2. Matchings rationalizable only by using thresholds.

Across the 39 schools, the median threshold level required to achieve rationalizability is 1 for rationalizability (no two connected minimal cycles), and 2 for TU/Extremal/Unique-rationalizability (no minimal cycles). In percentage scale, the median thresholds are 4.16% for rationalizability, and 4.76% for TU/Extreme/Unique-rationalizability.

As a result, we see that *the observed data do not distinguish much between matchings with and without transfers*. The reason is that romantic relationships tend to blossom between demographically similar individuals; mathematically, the resulting matching matrices are concentrated along the diagonal. Therefore, once we add enough noise in the model to generate a stable aggregate matching (no two connected cycles), that matching also tends to be compatible with stability and transfers (no cycles). Thus a stylized fact emerging from these results is that empirically, matchings with and without transfers may be difficult to distinguish. Of course, more careful empirical research in other settings is needed to establish these points more generally.

<sup>12</sup>This approach is common in the literature on revealed preference: see Varian (1985) or Echenique, Lee, and Shum (2011).

## 8. DISCUSSION

**8.1. Application to random matchings.** Our theory is also easily applicable to random matchings. For example consider a random matching where there is a set  $S$  of students and  $C$  of schools. Suppose that there is a set of two students  $S_0$  which have a positive chance of being admitted at every school. Then the resulting matching cannot be student optimal, no matter what the students' preferences are, or what schools preferences (priorities) are chosen. To see this simple point, imagine that students are men and schools are women in the notation above. Then if we consider the rows corresponding to the students in  $S_0$  we will have only positive entries:

$$\begin{pmatrix} & \vdots & \vdots & \\ \cdots & p_{s,c} & p_{s,c'} & \cdots \\ \cdots & p_{s',c} & p_{s',c'} & \cdots \\ & \vdots & \vdots & \end{pmatrix}$$

The graph  $(V, L)$  would then contain the following cycle as a subgraph.

$$\begin{array}{ccc} p_{s,c} & \text{---} & p_{s,c'} \\ | & & | \\ p_{s',c} & \text{---} & p_{s',c'} \end{array}$$

By Theorem 3, this random matching is not rationalizable as either student optimal or school optimal, regardless of what preferences one may assume for the two sides.

**8.2. Strong Stability.** Our definition of stability corresponds to *strong stability* for aggregate matchings, and *ex-ante stability* for random matchings. In this section we argue that (a) strong stability for aggregate matchings, and ex-ante stability for random matchings are natural notions of stability; and (b) that one cannot analyze strong stability using the standard linear programming tools.

First, it is rather obvious that strong stability is the natural notion of stability for aggregate matchings. Concretely, if there is a pair  $(m, w)$  that can block an aggregate matching  $X$ , then there are a man of type  $m$  and a woman of type  $w$ , such that the two agents can block in the usual sense. However, the case of random matching requires more explaining.

If  $X$  is a random matching, we can view the row  $X_m$  as a probability distribution over the set of women who may be partners of  $m$ ; similarly for  $X_w$ . Then  $(m, w)$  is a blocking pair (in the strong sense) if and only if there are distributions  $X'_m$  and  $X'_w$  such that

- $X'_m$  first order stochastically dominates  $X_m$ , and  $X'_w$  first order stochastically dominates  $X_w$ ;

- $(m, w)$  can achieve  $X'_m$  and  $X'_w$  by mutual agreement, without the consent of any other agents, because  $x'_{m,\tilde{w}} \leq x_{m,\tilde{w}}$  and  $x'_{\tilde{m},w} \leq x_{\tilde{m},w}$  for any  $\tilde{m} \neq m$  and  $\tilde{w} \neq w$ .

The ex-ante perspective makes sense if we think that agents can trade probabilities or time shares — see Hylland and Zeckhauser (1979). After agents trade, any random matching can be implemented physically by a decomposition into deterministic, simultaneous, matchings (using the Birkhoff von-Neumann theorem). If trades in probabilities or time-shares are somehow ruled out, then one can still justify strong stability on fairness grounds (Kesten and Ünver, 2009).

The standard notion of stability can be analyzed by linear programming methods (Vande Vate, 1989; Rothblum, 1992; Roth, Rothblum, and Vande Vate, 1993). The reason is that stability results in a collection of linear constraints on matchings. For instance, ‘no-blocking pair’ condition is represented as

$$\text{for all } (m, w), \quad \sum_{w':wP_m w'} x_{m,w'} + \sum_{m':mP_w m'} x_{m',w} \leq 1.$$

A fractional matching  $X$  is stable if and only if it satisfies a set of linear system of inequalities.

On the other hand, strong stability does not result in linear constraints. The property that an individually rational  $X$  is strongly stable is equivalent to the following statement. For all pairs  $(m, w)$ ,

$$\left( \sum_{w':wP_m w'} x_{m,w'} \right) \left( \sum_{m':mP_w m'} x_{m',w} \right) = 0.$$

Hence  $X$  is a strongly stable matching if and only if it is feasible, individually rational, and satisfies a set of quadratic constraints. The resulting set of matrices does not have the geometry that one can exploit for fractional stable matchings.

## 9. CONCLUSION

We study the testable implications of stability in two-sided matching markets. We present simple nonparametric tests for stability, extremal stability, and stability with monetary transfers. The tests are readily applicable to data on aggregate matchings and probabilistic matchings. The tests also have some surprising consequences: the theory of stable matching encompasses empirically the theory of stable matching with transfers; the latter is in turn observationally equivalent to the theory of extremal stable matching.

Our tests are easily applicable to the kinds of data used by empirical researchers. We present an illustration using high school romantic relationships from the Add Health data. The application, while brief, illustrates how the methods presented in this paper can be used to generate substantive stylized facts about empirical matching data.

## APPENDIX A. PROOF OF THEOREM 1

We split up the proof in short propositions.

**Lemma 1.**  $(\mathcal{X}_m, \leq_m)$  is a complete lattice, with a largest and smallest element.

*Proof.* That  $(\mathcal{X}_m, \leq_m)$  is a partially ordered set follows from the definition of  $\leq_m$ . Take a subset  $S_m$  of  $X_m$ . We need to show that  $S_m$  has a least upper bound and a greatest lower bound in  $(\mathcal{X}_m, \leq_m)$  to complete the proof.

For every  $x \in \mathcal{X}_m$ , define  $x_0 = K_m - \sum_{1 \leq j \leq |W|} x_j$ . By definition, for every  $x \in \mathcal{X}_m$ ,  $\sum_{0 \leq j \leq |W|} x_j = K_m$ . Assign a new index  $\{(j) : j = 0, 1, 2, \dots, |W|\}$  over  $W \cup \{w_0\}$  such that

$$w_{(0)}P_m w_{(1)}P_m \dots P_m w_{(|W|)}$$

Let  $z_{(0)} = \sup\{x_{(0)} : x \in S_m\}$ . Define  $z_{(i)}$  inductively as follows:  $z_{(i)} = \sup\{x_{(0)} + \dots + x_{(i)} : x \in S_m\} - (z_{(0)} + \dots + z_{(i-1)})$ .

Note that  $z_{(j)} \geq 0$  and  $\sum_{0 \leq j \leq |W|} z_{(j)} = K_m$ . We define  $z$  as a  $|W|$ -vector,  $(z_1, \dots, z_{|W|})$ , with respect to the original index. Clearly,  $z_j \geq 0$  and  $\sum_{1 \leq j \leq |W|} z_j \leq K_m$ , so  $z \in \mathcal{X}_m$ .

**Claim 1.**  $z$  is a least upper bound of  $S_m$ .

*Proof.* By definition,  $\sum_{j=0}^i z_{(j)} = \sup\{x_{(0)} + \dots + x_{(i)} : x \in S_m\}$  which is greater than  $x_{(0)} + \dots + x_{(i)}$  for all  $x \in S_m$  and  $0 \leq i \leq |W|$ . Therefore,  $z \geq_m x$  for all  $x \in S_m$ , which means that  $z$  is an upper bound.

Suppose that  $z'$  is another upper bound of  $S_m$ . Therefore,  $z'_{(0)} + \dots + z'_{(i)} \geq x_{(0)} + \dots + x_{(i)}$  for all  $x \in S_m$  and  $0 \leq i \leq |W|$ . If we take the supremum of the right hand side, then we get  $z'_{(0)} + \dots + z'_{(i)} \geq \sup\{x_{(0)} + \dots + x_{(i)} : x \in S_m\}$  for all  $0 \leq i \leq |W|$ . On the other hand,  $z_{(0)} + \dots + z_{(i)} = \sup\{x_{(0)} + \dots + x_{(i)} : x \in S_m\}$ . The last two impressions imply  $z'_{(0)} + \dots + z'_{(i)} \geq z_{(0)} + \dots + z_{(i)}$  for all  $i$ , so  $z' \geq_m z$ . Thus,  $z$  is a least upper bound.  $\square$

Similarly we can construct a greatest lower bound as follows:  $u_{(0)} = \inf\{x_{(0)} : x \in S_m\}$ . Define  $u_{(i)}$  inductively:  $u_{(i)} = \inf\{x_{(0)} + \dots + x_{(i)} : x \in S_m\} - (u_{(0)} + \dots + u_{(i-1)})$ . The proof that  $u$  is a greatest lower bound is similar to the proof that  $z$  is a least upper bound, so it is omitted.  $\square$

For each  $m$ , let the choice  $C_m$  be defined as follows. For a vector  $x \in \mathbf{R}_+^{|W|}$ , let  $C_m(x)$  be the vector in

$$\{y \in \mathcal{X}_m : y_j \leq x_j, j = 1, \dots, |W|\}$$

that is maximal for  $\leq_m$ . In other words, if  $x$  represents the quantities of type of women available for  $m$ ,  $C_m$  chooses according to  $P_m$  from the best choice downwards until filling

quota  $K_m$ . Note that if  $w_0 P_m w_j$  then  $y_j = 0$ , and that  $y_0 = K_m - \sum_{j:w_j P_m w_0} y_j$ . Define  $C_w$  analogously.

**Proposition 4.**  $(S(M, W, P, K), \leq_M)$  is a nonempty and complete lattice.

*Proof.* A man pre-matching is a matrix  $A = (a_{m,w})_{M \times W}$  such that  $a_{m,w} \in \mathbf{R}_+$  and  $\sum_w a_{m,w} \leq K_m$ . A woman pre-matching is a matrix  $B = (b_{m,w})_{M \times W}$  such that  $b_{m,w} \in \mathbf{R}_+$  and  $\sum_m b_{m,w} \leq K_w$ .

We consider pairs  $(A, B)$ , where  $A$  is a man pre-matching, and  $B$  is a woman pre-matching, ordered by a partial order  $\leq$ . The order  $\leq$  is defined as  $(A, B) \leq (A', B')$  if

$$\forall m, \forall w, (A_m \leq_m A'_m \text{ and } B'_w \leq_w B_w).$$

The order  $\leq$  is a product order of complete lattices by Lemma 1, so that the set of all pairs  $(A, B)$  ordered by  $\leq$  is a complete lattice.

We define a function  $C$ , mapping pairs  $(A, B)$  of pre-matchings into pairs of pre-matchings. Fix  $(A, B)$ : For a man of type  $m$ , the number of women of type  $w$  who are willing to match with  $m$  at  $B$  is  $\theta_{m,w} = \sum_{i:mR_w m_i} b_{i,w}$ . Let  $\Theta = (\theta_{m,w})$ , i.e., the  $|M| \times |W|$ -matrix such that entry  $\theta_{m,w}$  is the number of type  $w$  women who are willing to match with type  $m$  men at  $B$ . Similarly, let  $\Psi = (\psi_{m,w})$  be the  $|M| \times |W|$ -matrix for which entry  $\psi_{m,w}$  is the number of type  $m$  men who are willing to match with type  $w$  women at  $A$ . Now let  $C(A, B) = (A', B')$  where  $A'_m = C_m(\Theta_m)$  and  $B'_w = C_w(\Psi_w)$ .

We now prove that  $C$  is isotone. Assuming that  $(A, B) \leq (A', B')$ , we prove that  $C(A, B) \leq C(A', B')$ .

Take any  $m$ . For any  $w$ , we obtain

$$\sum_{i:mR_w m_i} b_{i,w} = K_w - \sum_{i:m_i P_w m} b_{i,w} \leq K_w - \sum_{i:m_i P_w m} b'_{i,w} = \sum_{i:mR_w m_i} b'_{i,w},$$

because  $B'_w \leq_w B_w$ .

As a consequence, a type  $m$  man has weakly more women of each type willing to match with him in  $B'$  than in  $B$ : i.e.  $\Theta_m \leq \Theta'_m$ . Thus  $C_m(\Theta_m) \leq_m C_m(\Theta'_m)$ . Similarly, for a type  $w$  woman,  $C_w(\Psi'_w) \leq_w C_w(\Psi_w)$ . It follows that  $C(A, B) \leq C(A', B')$ . By Tarski's fixed point theorem, there is a fixed point of  $C$ , and the set of fixed points of  $C$  is a complete lattice when ordered by  $\leq$ .

Let  $(A, B) = C(A, B)$  be a fixed point of  $C$ . Assume that  $a_{m,w} > b_{m,w}$  for some  $m$  and  $w$ .  $C_w(\Psi_w) = B_w$  and  $a_{m,w} > b_{m,w}$  implies that although  $a_{m,w}$  number of type  $m$  men were available, only  $b_{m,w}$  of them are chosen by  $C_w$ . Therefore, all nonnegative entries in  $B_w$  are at least as good as  $m$  with respect to  $R_w$ . This implies that  $\theta_{m,w} = b_{m,w}$ . Since  $b_{m,w} < a_{m,w}$ , we get  $\theta_{m,w} < a_{m,w}$  which contradicts  $A_m = C_m(\Theta_m)$ . Therefore,  $a_{m,w} = b_{m,w}$  for all  $m$  and  $w$ . Hence a fixed point has the property that  $A = B$ , and they are not only a pre-matching but a matching as well.

Finally we prove that the set of fixed points of  $C$  is the set of stable matchings. More precisely,  $(A, A)$  is a fixed point of  $C$  if and only if  $A$  is a stable matching.

Suppose that a fixed point  $(A, A)$  is not stable. Then there is a blocking pair  $(m, w)$ . That is, there is  $m'$  and  $w'$  such that  $m P_w m'$ ,  $w P_m w'$ ,  $a_{m,w'} > 0$ , and  $a_{m',w} > 0$ . Now, the number of women of type  $w$  who are willing to match with  $m$  is

$$\theta_{m,w} = \sum_{i: m R_w m_i} a_{i,w} \geq a_{m,w} + a_{m',w} > a_{m,w},$$

as  $a_{m',w} > 0$ . But  $\theta_{m,w} > a_{m,w}$  and  $a_m = C_m(\Theta_m)$  contradicts that there is  $w'$  with  $w P_m w'$  and  $a_{m,w'} > 0$ .

Suppose that  $A$  is a stable matching. We fix  $m$  and show that  $a_m = C_m(\Theta_m)$  where  $\theta_{m,w} = \sum_{i: m R_w m_i} a_{i,w}$ . Denote  $w_j$  as the most preferred type of women (with respect to  $R_m$ ) such that  $a_{m,j} \neq (C_m(\Theta_m))_j$ . By definition of  $C_m$ ,  $a_{m,j} > (C_m(\Theta_m))_j$  is not feasible. For all  $w_{j'}$  preferred to  $w_j$ ,  $a_{m,j'} = (C_m(\Theta_m))_{j'}$ . Thus,  $a_{m,j} > (C_m(\Theta_m))_j$  implies either  $a_{m,j} > \theta_{m,j}$  or  $\sum_{j': w_{j'} R_m w_j} a_{m,j'} > K_m$ , neither of which are possible. On the other hand,  $a_{m,j} < (C_m(\Theta_m))_j$  contradicts that  $A$  is stable. Although there are type  $w_j$  women available more than  $a_{m,j}$ , some type  $m$  men are matched to less preferred women. Then, there is  $j'$  such that  $w_j P_m w_{j'}$  and  $a_{m,j'} > 0$ , so  $(m, w_j)$  is a blocking pair. Similarly, we can show that  $a_w = C_w(\Theta_w)$ , and therefore  $(A, A) = C(A, A)$ .  $\square$

**Proposition 5.** *Suppose that  $X$  and  $Y$  are stable matchings, then  $X \leq_M Y$  if and only if  $Y \leq_W X$ .*

*Proof.* Suppose that  $X$  and  $Y$  are stable matchings such that  $X \leq_M Y$ , we are going to show that  $Y \leq_W X$ . The other direction of the claim can be proved analogously.

Consider the construction of  $C$  in the proof of Proposition 4: Let  $\Psi$  and  $\Psi'$  correspond to the matrices for  $X$  and  $Y$  as in the proof. Since  $X$  and  $Y$  are stable matchings we have  $C(X, X) = (X, X)$  and  $C(Y, Y) = (Y, Y)$ . Since  $X \leq_m Y$  for all  $m$ ,  $\psi'_{m,w} \leq \psi_{m,w}$  for all  $m$  and  $w$ . Therefore,  $Y_w = C_w(\psi'_w) \leq_w C_w(\psi_w) = X_w$  for all  $w$ , which implies that  $Y \leq_W X$ .  $\square$

**Proposition 6.** *Suppose that  $X$  and  $Y$  are two stable matchings. Then for any men or women type  $a$ , either  $X_a \leq_a Y_a$  or  $Y_a \leq_a X_a$ . Consequently,  $X_a \vee_a Y_a = \max_{\leq_a} \{X_a, Y_a\}$  and  $X_a \wedge_a Y_a = \min_{\leq_a} \{X_a, Y_a\}$ .*

*Proof.* We only prove the first part that either  $X_a \leq_a Y_a$  or  $Y_a \leq_a X_a$ , in three steps depending on whether  $X$  and  $Y$  have integer, rational, and real entries. The second part follows immediately.

Case 1 (Integer Entries): We first start with the case when  $X$  and  $Y$  have integer entries. Therefore,  $K_w$  and  $K_m$  are also integers. From  $\langle M, W, P, K \rangle$ , we create a many-to-one matching market (of colleges and students) as follows.

The set of types of men remains the same; interpreted as the set of colleges. A college  $m$  has a capacity of  $K_m$ . Whereas, each woman of type  $w$  is split into  $K_w$  copies, all of which have the same preferences  $P_w$  over colleges; women types are interpreted as students. On the other hand,  $m$ 's preferences  $P'_m$  replaces  $w$  in  $P_m$  with her copies enumerated from 1 to  $K_w$ , in increasing order. Denote  $w_j$ 's  $l^{\text{th}}$  copy by  $w_j^l$ . So  $w_j^l P'_m w_j^k$  if and only if  $k > l$ . In addition, each college has responsive preferences over groups of students. The new matching market with  $|M|$  colleges and  $\sum_w K_w$  students is a many-to-one matching market where an outcome for a college is a group of students, and an outcome for a student is either a college or being single.

Now, we construct a new matching,  $X'$ , in the new market from  $X$ .<sup>13</sup> It is enough to describe the matches of students in  $X'$ . Rank woman  $w$ 's outcomes in  $X$  in decreasing order according to her preference  $P_w$ . Let the  $l^{\text{th}}$  copy of  $w_j$ ,  $w_j^l$ , match to the  $l^{\text{th}}$  highest outcome of  $w_j$  in  $X$ . Similarly, construct  $Y'$  from  $Y$ .

We claim that  $X'$  and  $Y'$  are stable matchings in the new market. Suppose for contradiction that  $X'$  is not a stable matching. Since  $X'$  is individually rational by construction, there exists a blocking pair  $(m, w_j^l)$ . This means that  $w_j^l$ 's match is worse than  $m$ . Similarly,  $m$ 's match includes a student worse than  $w_j^l$ : this agent cannot be  $w_j^k$  where  $k < l$  by definition of  $P'_m$ , and it cannot be  $w_j^k$  where  $k > l$  because by construction  $w_j^k$ 's match is worse than  $w_j^l$ 's match with respect to  $P_{w_j}$ . Hence, one of  $m$ 's matches is worse than  $w_j^l$ , and not a copy of  $w_j$ . This means that  $(m, w_j)$  forms a blocking pair in  $X$ : A contradiction to the stability of  $X$ . Therefore,  $X'$  must be stable. Similarly,  $Y'$  is also stable.

Now, by Theorem 5.26 of Roth and Sotomayor (1990), for any college  $m$  the outcomes in  $X'$  and  $Y'$  are comparable. This means that the responsive preferences over groups of students inherited from  $P'_m$ , which is equivalent to the first order stochastic dominance, can compare the outcomes of  $m$  in these two stable matchings. Therefore,  $\leq_m$  can compare the outcomes in  $X$  and  $Y$  since  $P_m$  is a coarser order than  $P'_m$ .

An analogous argument shows that  $\leq_w$  can compare the matching outcomes in  $X$  and  $Y$ .

Case 2 (Rational Entries): Suppose for now that all entries of  $X$  and  $Y$  are rational numbers. Therefore,  $K_w$  and  $K_m$  are also rational numbers. Define a new matching market from  $\langle M, W, P, K \rangle$  as follows.

$M$ ,  $W$ , and  $P$  are the same. We change the capacities as follows. Find a common factor of denominators in all entries in  $X$  and  $Y$ , say  $r$ , and multiply all capacities by  $r$ . Therefore, the new market is  $\langle M, W, P, rK \rangle$ . Define  $X' = rX$  and  $Y' = rY$  with non-negative integer entries. By the argument above, for any agent  $a$ , the matchings in  $rX$  and  $rY$  can be compared with respect to  $\leq_a$  which implies that the outcomes in  $X$  and  $Y$  can also be compared.

<sup>13</sup>The new matching is a many-to-one matching, which is not an aggregate matching. Nevertheless, we use the aggregate matching notation to stress the relations between  $X'$  and  $X$ , and  $Y'$  and  $Y$ .

Case 3 (Real Entries): Suppose now that entries of  $X$  and  $Y$  are real numbers. We construct two sequences of matrices,  $X^{(n)}$  and  $Y^{(n)}$ , as follows:

- (1)  $X^{(n)}$  and  $Y^{(n)}$  have rational entries,
- (2)  $x_{ij}^{(n)} = 0 \iff x_{ij} = 0$  and  $y_{ij}^{(n)} = 0 \iff y_{ij} = 0$  for all  $i, j$ ,
- (3) the sum of entries in row  $i$  and column  $j$  is the same for  $X^{(n)}$  and  $Y^{(n)}$ , and
- (4)  $x_{ij}^{(n)} \rightarrow x_{ij}$  and  $y_{ij}^{(n)} \rightarrow y_{ij}$  as  $n \rightarrow \infty$  for all  $i, j$ .

By construction, stability of  $X$  and  $Y$  imply stability of  $X^{(n)}$  and  $Y^{(n)}$  for the same market with adjusted capacities. By the argument above, for each type  $a$ ,  $x_a^{(n)}$  and  $y_a^{(n)}$  can be compared with respect to  $\leq_a$ . Take a subsequence  $\{n_l\}$  such that the ordering is the same for all entries. Therefore,  $x_a^{(n_l)} \leq_a y_a^{(n_l)}$  for all  $l$  or  $y_a^{(n_l)} \leq_a x_a^{(n_l)}$  for all  $l$ . By taking  $l$  to  $\infty$  we get that either  $x_a \leq_a y_a$  in the former case, or  $y_a \leq_a x_a$  in the latter since (4) implies  $x_{i,j}^{(n)} \rightarrow x_{i,j}$  and  $y_{i,j}^{(n)} \rightarrow y_{i,j}$  as  $n \rightarrow \infty$  for all  $i, j$ .  $\square$

We have shown that  $(S(M, W, P, K), \leq_M)$  is a lattice in Proposition 4. Using the proposition above, we show that the lattice,  $(S(M, W, P, K), \leq_M)$ , is distributive.

**Proposition 7.**  $(S(M, W, P, K), \leq_M)$  is a distributive lattice.

*Proof.* Suppose that  $X, Y$ , and  $Z$  are stable matchings. We are going to prove that type  $a$  have the same matching in  $X \wedge (Y \vee Z)$  and  $(X \wedge Y) \vee (X \wedge Z)$  for all agent types  $a$ :

$$\begin{aligned} (X \wedge (Y \vee Z))_a &= \min\{X_a, \max\{Y_a, Z_a\}\} \\ &= \max\{\min\{X_a, Y_a\}, \min\{X_a, Z_a\}\} \\ &= \max\{(X \vee Y)_a, (X \vee Z)_a\} \\ &= ((X \wedge Y) \vee (X \wedge Z))_a, \end{aligned}$$

where min and max operators are defined with respect to  $\leq_a$  and where we repeatedly use Proposition 6.

The proof that  $X \vee (Y \wedge Z) = (X \vee Y) \wedge (X \vee Z)$  is analogous.  $\square$

**Proposition 8.** Suppose that  $X$  and  $Y$  are stable matchings. Then  $\sum_{1 \leq j \leq |W|} x_{ij} = \sum_{1 \leq j \leq |W|} y_{ij}$  for all  $i$  and similarly  $\sum_{1 \leq i \leq |M|} x_{ij} = \sum_{1 \leq i \leq |M|} y_{ij}$  for all  $j$ .

*Proof.* The proof has the same structure as in the proof of Proposition 6: it has three steps depending on whether  $X$  and  $Y$  have integer, rational, and real entries.

Case 1 (Integer Entries): We first start with the case when  $X$  and  $Y$  have integer entries. Therefore,  $K_a$  is also an integer for all  $a$ . From  $\langle M, W, P, K \rangle$ , we create a many-to-one matching market and also stable matchings  $X'$  and  $Y'$  in this new market as in the proof of Proposition 6.

Now, by Theorem 5.12 of (Roth and Sotomayor, 1990), the set of positions filled for any  $m$  in  $X'$  and  $Y'$  are the same. Therefore,  $\sum_{1 \leq j \leq |W|} x_{ij} = \sum_{1 \leq j \leq |W|} y_{ij}$  for all  $i$ . Similarly,

$$\sum_{1 \leq i \leq |M|} x_{ij} = \sum_{1 \leq i \leq |M|} y_{ij} \text{ for all } j.$$

Case 2 (Rational Entries): Suppose for now that all entries of  $X$  and  $Y$  are rational numbers. Therefore,  $K_a$  is also a rational number for all  $a$ . Define a new matching market from  $\langle M, W, P, K \rangle$  as follows.

$M$ ,  $W$ , and  $P$  are the same. We change the capacities as follows. Find a common factor of denominators in all entries in  $X$  and  $Y$ , say  $r$ , and multiply all capacities by  $r$ . Therefore, the new market is  $\langle M, W, P, rK \rangle$ . Define  $X' = rX$  and  $Y' = rY$  with non-negative integer entries.

By the argument above,  $\sum_{1 \leq j \leq |W|} rx_{ij} = \sum_{1 \leq j \leq |W|} ry_{ij}$  for all  $i$  and  $\sum_{1 \leq i \leq |M|} rx_{ij} = \sum_{1 \leq i \leq |M|} ry_{ij}$  for all  $j$ . The conclusion follows.

Case 3 (Real Entries): Suppose now that entries of  $X$  and  $Y$  are real numbers. We construct two sequences of matrices,  $X^{(n)}$  and  $Y^{(n)}$ , as follows:

- (1)  $X^{(n)}$  and  $Y^{(n)}$  have rational entries,
- (2)  $x_{ij}^{(n)} = 0 \iff x_{ij} = 0$  and  $y_{ij}^{(n)} = 0 \iff y_{ij} = 0$  for all  $i, j$ ,
- (3) the sum of entries in row  $i$  and column  $j$  is the same for  $X^{(n)}$  and  $Y^{(n)}$ , and
- (4)  $x_{ij}^{(n)} \rightarrow x_{ij}$  and  $y_{ij}^{(n)} \rightarrow y_{ij}$  as  $n \rightarrow \infty$  for all  $i, j$ .

By construction, stability of  $X$  and  $Y$  imply stability of  $X^{(n)}$  and  $Y^{(n)}$  for the same market with adjusted capacities. By the argument above,  $\sum_{1 \leq j \leq |W|} x_{ij}^{(n)} = \sum_{1 \leq j \leq |W|} y_{ij}^{(n)}$  for all  $i$  and

$\sum_{1 \leq i \leq |M|} x_{ij}^{(n)} = \sum_{1 \leq i \leq |M|} y_{ij}^{(n)}$ . If we take the limit of these equalities as  $n \rightarrow \infty$ , we get the desired equalities.  $\square$

## APPENDIX B. PROOF OF THEOREM 2.

**B.1. Proof of necessity.** We break up the proof into a collection of lemmas.

First, note a simple fact about minimal cycles:

**Lemma 2.** *If  $c = \langle v_0, \dots, v_N \rangle$  is a minimal cycle, then no vertex appears twice in  $c$ .*

An *orientation* of  $(V, L)$  is a mapping  $d : L \rightarrow \{0, 1\}$ . We shall often write  $d((m, w), (m, w'))$  as  $d_{m,w,w'}$  and  $d((m, w), (m', w))$  as  $d_{w,m,m'}$ . A preference profile  $P$  defines an orientation  $d$  by setting  $d_{w,m,m'} = 1$  if and only if  $m P_w m'$ , and  $d_{m,w,w'} = 1$  if and only if  $w P_m w'$ .

Let  $d$  be an orientation defined from a preference profile. Then  $X$  is stable if and only if, for all  $(m_1, w_1)$  and  $(m_2, w_2)$ , if  $x_{1,1} > 0$  and  $x_{2,2} > 0$  then

$$(3) \quad d_{m_1, w_2, w_1} d_{w_2, m_1, m_2} = 0 \text{ and } d_{m_2, w_1, w_2} d_{w_1, m_2, m_1} = 0.$$

We say that the pair  $((m_1, w_1), (m_2, w_2))$  is an *antiedge* if  $x_{1,1} > 0$  and  $x_{2,2} > 0$  for  $m_1 \neq m_2$  and  $w_1 \neq w_2$ .

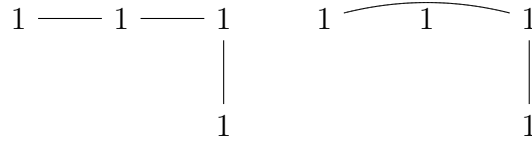
Fix an orientation  $d$  of  $(V, L)$ . A path  $\{(m, w)_n : n = 0, \dots, N\}$  is a *flow* for  $d$  if either  $d((m, w)_n, (m, w)_{n+1}) = 1$  for all  $n \in \{0, \dots, N-1\}$ , or  $d((m, w)_n, (m, w)_{n+1}) = 0$  for all  $n \in \{0, \dots, N-1\}$ . If the second statement is true, we call the path a *forward flow*.

Our first observation is an obvious consequence of the property of being minimal:

**Lemma 3.** *Let  $\{(m, w)_n : n = 0, \dots, N\}$  be a minimal path with  $N \geq 2$ , then for any  $n \in \{0, \dots, N-2\}$ ,*

$$(m_n = m_{n+1} \Rightarrow w_{n+1} = w_{n+2}) \text{ and } (w_n = w_{n+1} \Rightarrow m_{n+1} = m_{n+2}).$$

That is, any two subsequent edges in a path must be at a right angle:



The path on the left is not minimal; the path on the right is.

Fix an orientation  $d$  derived from the preferences rationalizing  $X$ .

**Lemma 4.** *Let  $p = \langle (m, w)_n : n = 0, \dots, N \rangle$  be a minimal path. If  $d((m, w)_1, (m, w)_0) = 1$  or  $d((m, w)_N, (m, w)_{N-1}) = 0$ , then  $p$  is a flow for  $d$ .*

*Proof.* By Lemma 3, for any  $n \in \{1, \dots, N-1\}$  the pair of vertices  $(m, w)_{n-1}$  and  $(m, w)_{n+1}$  form an antiedge: we have  $x_{(m, w)_{n-1}} > 0$ ,  $x_{(m, w)_{n+1}} > 0$ ,  $m_{n-1} \neq m_{n+1}$ , and  $w_{n-1} \neq w_{n+1}$ . Further,  $(m, w)_n$  has one element in common with  $(m, w)_{n-1}$  and the other in common with  $(m, w)_{n+1}$ . Thus by Equation 3,  $d((m, w)_n, (m, w)_{n-1}) = 1$  implies that  $d((m, w)_n, (m, w)_{n+1}) = 0$ : i.e.  $d((m, w)_{n+1}, (m, w)_n) = 1$ .

The argument in the previous paragraph shows that the existence of some  $n'$  with  $d((m, w)_{n'}, (m, w)_{n'-1}) = 1$  implies  $d((m, w)_n, (m, w)_{n-1}) = 1$  for all  $n \geq n'$ . So if  $d((m, w)_1, (m, w)_0) = 1$  then  $d((m, w)_{n+1}, (m, w)_n) = 1$  for all  $n \in \{1, \dots, N-1\}$ ; and if  $d((m, w)_N, (m, w)_{N-1}) = 0$ , then  $d((m, w)_{n+1}, (m, w)_n) = 0$  for all  $n \in \{0, \dots, N-1\}$ . Either way,  $p$  is a flow.  $\square$

As an immediate consequence of Lemma 4, we obtain the following

**Lemma 5.** *Let  $p = \langle (m, w)_n \rangle$  be a minimal cycle, then  $p$  is a flow for  $d$ .*

Let  $p = \langle (m, w)_n \rangle$  be a path and  $(m, w) \notin p$ . A path  $\bar{p} = \langle (\bar{m}, \bar{w})_n : n = 0, \dots, \bar{N} \rangle$  connects  $p$  and  $(m, w)$  if  $(\bar{m}, \bar{w})_0 \in p$  and  $(\bar{m}, \bar{w})_{\bar{N}} = (m, w)$ .

**Lemma 6.** *Let  $c = \langle (m, w)_n \rangle$  be a minimal cycle, and  $p = \langle (\bar{m}, \bar{w})_n : n = 0, \dots, \bar{N} \rangle$  be a minimal path connecting  $c$  to some  $(\bar{m}, \bar{w})$ . Then  $\langle (\bar{m}, \bar{w})_n : n = 1, \dots, \bar{N} \rangle$  is a forward flow.*



$N < 3$  we can add  $(m', w') \in c_1$  to  $p$  with  $((m', w'), (m, w)_0) \in L$ , and  $(m'', w'') \in c_2$  to  $p$  with  $((m'', w''), (m, w)_N) \in L$ ; the corresponding path will also be a minimal path connecting  $c_1$  and  $c_2$ .

By Lemma 6 applied to  $c_1$  and  $p$ ,

$$\langle (m, w)_n : n = 1, \dots, N \rangle$$

is a forward flow. On the other hand, Lemma 6 applied to  $c_2$  and  $p$  implies that

$$\langle (m, w)_{N-k} : k = 1, \dots, N \rangle$$

is a forward flow. The first statement implies that  $d((m, w)_2, (m, w)_1) = 1$  and the second that  $d((m, w)_1, (m, w)_2) = 1$ , a contradiction.  $\square$

**B.2. Proof of sufficiency.** To prove sufficiency, we explicitly construct an orientation  $d$  that satisfies Equation (3). We then show that there is a rationalizing preference profile. In Appendix F.1, we construct a rationalizing preference profile for an example.

We first deal with the case where all vertices in  $X$  are connected and there is at most one minimal cycle. By decomposing an arbitrary  $X$  into connected components, we shall later generalize the argument. If there is no cycle in  $X$ , choose a singleton vertex and treat it as the “cycle” in the sequel.

Let  $C$  be the submatrix having the indices in the minimal cycle. If  $c = \langle (m, w)_n \rangle$  is the minimal cycle, let  $M_1 = \cup_n \{m_n\}$  and  $W_1 = \cup_n \{w_n\}$ . Then  $C$  is the matrix  $(x_{m', w'})_{(m', w') \in M_1 \times W_1}$ . Thus  $C$  contains the minimal cycle.

We re-arrange the indices of  $X$  to obtain a matrix of the form:

$$\begin{array}{c|ccc} & (W_1) & (W_2) & (W_3) \\ \hline (M_1) & C & X_1 & O & \cdots \\ (M_2) & Y_1 & O & X_2 & \cdots \\ (M_3) & O & Y_2 & O & \cdots \\ & \vdots & \vdots & \vdots & \end{array}$$

We define the submatrices  $X_n$  and  $Y_n$  by induction. For  $n \geq 1$ , let

$$M_{n+1} = \{m \notin \cup_1^n M_k : \exists w \in \cup_1^n W_k \text{ s.t. } (m, w) \in V\},$$

$$W_{n+1} = \{w \notin \cup_1^n W_k : \exists m \in \cup_1^n M_k \text{ s.t. } (m, w) \in V\}.$$

Now, let  $X_n$  be the matrix  $(x_{m', w'})_{(m', w') \in M_n \times W_{n+1}}$  and  $Y_n$  be the matrix  $(x_{m', w'})_{(m', w') \in M_{n+1} \times W_n}$ . Finally, re-label the indices such that if  $m_i \in M_n$  and  $m_{i'} \in M_{n'}$  and  $n < n'$  then  $i < i'$ . The numbering of indexes in  $M_n$  is otherwise arbitrary. Re-label  $w$ 's in a similar fashion.

For every  $m \in M_n$  there is a  $k < n$  and  $w \in W_k$  such that  $(m, w) \in V$ , and similarly, for every  $w \in W_n$  there is a  $k < n$  and  $m \in M_k$  such that  $(m, w) \in V$ . Thus, for  $m \in M_n$  there

is a sequence

$$(m, w_{k_0}), (m_{k_1}, w_{k_0}), \dots, (m_{k_N}, w_{k_{N'}}),$$

with  $N = N' + 1$  or  $N' = N - 1$ , which defines a path connecting  $(m, w_{k_0})$  to the cycle  $c$ . Similarly, if  $w \in W_n$  there is a path connecting  $(m_{k_0}, w)$  to  $c$ .

The observation in the previous paragraph has two consequences:

**Claim 2.** *If  $m \in M_n$  and  $w \in W_n$  ( $n > 1$ ), then  $(m, w) \notin V$ .*

Claim 2 is true because otherwise there would be two different paths connecting  $(m, w)$  to  $c$ , one having  $(m, w_{k_0})$  and the other  $(m_{k_0}, w)$  as second element. Then we would have a distinct second cycle.

**Claim 3.** *Let  $m_i \in M_n$  ( $n > 1$ ), and let there be two distinct  $w_j$  and  $w_{j'}$  ( $j' > j$ ) such that  $(m_i, w_j), (m_i, w_{j'}) \in V$ . Then  $(m_{i'}, w_{j'}) \in V$  implies that  $m_{i'} \in M_{n'}$  with  $n' > n$ .*

Claim 3 is true because otherwise we would again have two different paths connecting  $(m_i, w_{j'})$  to  $c$ ; one path with  $(m_i, w_j)$  and one with  $(m_{i'}, w_{j'})$  as its second element.

Define the orientation  $d$  as follows. We simply use  $d_{i,j,j'}$  for  $d_{m_i, w_j, w_{j'}}$ , and  $d_{j,i,i'}$  for  $d_{w_j, m_i, m_{i'}}$ .

- (1) If  $(m_i, w_j) \in c$  and  $(m_i, w_{j'}) \in c$  then define  $d_{i,j,j'}$  to be 1 if  $(m_i, w_j)$  comes immediately after  $(m_i, w_{j'})$  in  $c$ . That is,  $d_{i,j,j'} = 1$  if there is  $n$  such that

$$(m_i, w_{j'}) = (m_i, w_j)_n \text{ and } (m_i, w_j) = (m_i, w_{j'})_{n+1}.$$

- (2) If  $(m_i, w_j) \in c$  and  $(m_{i'}, w_j) \in c$  then define  $d_{j,i,i'}$  to be 1 if  $(m_i, w_j)$  comes immediately after  $(m_{i'}, w_j)$  in  $c$ .
- (3) If  $(m_i, w_j) \notin c$  and  $(m_i, w_{j'}) \in c$  then define  $d_{i,j,j'}$  to be 1.
- (4) If  $(m_i, w_j) \notin c$  and  $(m_{i'}, w_j) \in c$  then define  $d_{j,i,i'}$  to be 1.
- (5) If  $(m_i, w_j) \notin c$  and  $(m_i, w_{j'}) \notin c$  then define  $d_{i,j,j'}$  to be 1 if and only if  $j > j'$ .
- (6) If  $(m_i, w_j) \notin c$  and  $(m_{i'}, w_j) \notin c$  then define  $d_{j,i,i'}$  to be 1 if and only if  $i > i'$ .
- (7) If  $(m_i, w_j) \in V$  and  $(m_{i'}, w_j) \notin V$ , then define  $d_{j,i,i'}$  to be 1.

Let  $d_{i,j',j} = 0$  when 1-7 imply that  $d_{i,j,j'} = 1$ ; similarly  $d_{j,i',i} = 0$  when 1-7 imply that  $d_{j,i,i'} = 1$ .

**Lemma 8.** *If  $(m_i, w_j)$  is a vertex in  $c$ , then there is at most one  $w_{j'}$  such that  $j' \neq j$  and  $(m_i, w_{j'}) \in c$ ; in addition,  $(m_i, w_j)$  and  $(m_i, w_{j'})$  are adjacent in  $c$ . Similarly, there is at most one  $i' \neq i$  such that  $(m_{i'}, w_j) \in c$ ; in addition,  $(m_i, w_j)$  and  $(m_{i'}, w_j)$  are adjacent in  $c$ .*

*Proof.* We let the index of  $c$  range over all the integers by denoting  $(m, w)_n \pmod{N}$  by  $(m, w)_n$ .

Let  $(m, w)$  be a vertex in  $c$ , and  $n > 0$  be such that  $(m, w) = (m, w)_n$ . Suppose there is  $w'$  such that  $w' \neq w$  and  $(m, w') \in c$ . If it does not exist, we are done. Since now  $N \geq 2$ ,

$(m, w)$  is in the minimal path connecting  $(m, w)_{n-1}$  and  $(m, w)_{n+1}$ . By Lemma 3, then, either  $m_{n-1} = m$  or  $m_{n+1} = m$ , and exactly one of these is true. In the first case, we can set  $w' = w_{n-1}$  and in the second we can set  $w' = w_{n+1}$ . Suppose, without loss of generality, that  $w' = w_{n+1}$ .

We show that there is not a  $w'' \neq w, w'$  with  $(m, w'') \in c$ . Suppose that there is such a  $w''$ . Let  $(m, w'') = (m, w)_l$ . By Lemma 3, we have either  $l < n - 1$  or  $l > n + 1$ . When  $l > n + 1$ , the path  $\langle (m, w)_{n-1}, \dots, (m, w)_l \rangle$  is not minimal because  $\langle (m, w)_{n-1}, (m, w)_n, (m, w)_m \rangle$  is a proper subset connecting  $(m, w)_{n-1}$  and  $(m, w)_m$ . When  $l < n - 1$ , the path  $\langle (m, w)_m, (m, w)_n, (m, w)_{n+1} \rangle$  is not a minimal because  $(m, w)_m$  and  $(m, w)_{n+1}$  are directly connected. Thus  $c$  is not a minimal cycle, a contradiction.  $\square$

**Lemma 9.** *Let  $(m, w)$  be a vertex in  $c$ . If  $(m, w') \in V$  is not a vertex in  $c$ , then, for all  $m' \neq m$ ,  $(m', w') \notin c$ . Similarly, if  $(m', w) \in V$  is not a vertex in  $c$ , then, for all  $w' \neq w$ ,  $(m', w') \notin c$ .*

*Proof.* Suppose for contradiction that  $(m, w) \in c$ ,  $(m', w') \in c$ , with  $m \neq m'$ ,  $w \neq w'$ , and  $(m, w') \notin c$ . Since  $(m, w), (m', w') \in c$ , there is a minimal path  $\langle (m, w)_n : n = 0, \dots, N \rangle$  connecting  $(m', w')$  to  $(m, w)$ . Then, since  $(m, w') \notin c$ , the minimal cycle

$$\langle (m, w)_0, \dots, (m, w)_N, (m, w'), (m', w') \rangle$$

is distinct from  $c$  and connected to  $c$ .  $\square$

**Lemma 10.** (1) *If  $d_{i,j,j'} = 1$  and  $d_{i,j',j''} = 1$ , then  $d_{i,j,j''} = 1$ .*

(2) *If  $d_{j,i,i'} = 1$  and  $d_{j,i',i''} = 1$ , then  $d_{j,i,i''} = 1$ .*

*Proof.* We prove only the first statement. The second statement can be proved by similar fashion to the following first three cases.

First, we can rule out that  $d_{i,j,j'} = 1$  because  $(m_i, w_j) \in c$ ,  $(m_i, w_{j'}) \in c$ , and  $(m_i, w_j)$  comes immediately after  $(m_i, w_{j'})$  in  $c$  (Case 1). To see this, note that  $d_{i,j',j''} = 1$  would imply that either  $(m_i, w_{j''}) \in c$ , which is not possible by Lemma 8.

Second, suppose that  $d_{i,j,j'} = 1$  is from the face that  $(m_i, w_j) \notin c$  and  $(m_i, w_{j'}) \in c$ . Then  $d_{i,j',j''} = 1$  implies that  $(m_i, w_{j''}) \in c$ . Thus  $d_{i,j,j''} = 1$  by Case 3.

Third, suppose that  $d_{i,j,j'} = 1$  is from the fact that  $(m_i, w_j) \notin c$ ,  $(m_i, w_{j'}) \notin c$ , and  $j > j'$ . If  $d_{i,j',j''} = 1$  because  $(m_i, w_{j''}) \notin c$  and  $j' > j''$ , then  $d_{i,j,j''} = 1$  by Case 5 and the transitivity of  $P$ . On the other hand, if  $d_{i,j',j''} = 1$  because  $(m_i, w_{j''}) \in c$ , then  $d_{i,j,j''} = 1$  (Case 3) as well. Finally, if  $d_{i,j,j'} = 1$  because of Case 7 then we obtain  $d_{i,j,j''} = 1$  by Case 7 as well.  $\square$

**Lemma 11.** *The orientation  $d$  satisfies (3).*

*Proof.* Let  $((m_i, w_j), (m_{i'}, w_{j'}))$  be an antiedge: so  $(m_i, w_j), (m_{i'}, w_{j'}) \in V$ ,  $j \neq j'$  and  $i \neq i'$ . Suppose that  $d_{i,j',j} = 1$ . We shall prove that  $d_{j',i,i'} = 0$ .

Suppose first that  $d_{i,j',j} = 1$  because of Case 1. Then  $(m_i, w_j) \in c$ . So, if  $(m_{i'}, w_{j'}) \notin c$  we obtain that  $d_{j',i,i'} = 0$  by Case 3. On the other hand, if  $(m_{i'}, w_{j'}) \in c$  then the edges  $((m_i, w_j), (m_i, w_{j'}))$  and  $((m_i, w_{j'}), (m_{i'}, w_{j'}))$  are in  $c$ . In fact, these edges must be consecutive, or  $(m_i, w_{j'})$  will appear twice in  $c$ . Then,  $d_{i,j',j} = 1$  because of Case 1 implies that  $(m_i, w_{j'})$  comes immediately after  $(m_i, w_j)$  in  $c$ ; the edge  $((m_i, w_{j'}), (m_{i'}, w_{j'}))$  comes after  $((m_i, w_j), (m_i, w_{j'}))$  in  $c$ , so we obtain that  $d_{j',i,i'} = 0$  by Case 1.

Suppose second that  $d_{i,j',j} = 1$  because of Case 3. So  $(m_i, w_j) \in c$  and  $(m_i, w_{j'}) \notin c$ . Then  $m_i \in M_1$  because  $m_i$  is an index for a vertex in the minimal cycle  $c$ . Now, by Lemma 9, there is no  $\tilde{m}_i$  with  $(\tilde{m}_i, w_{j'}) \in c$ . Since  $(m_{i'}, w_{j'}) \in V$  we must have  $m_{i'} \in M_n$  for  $n > 1$ . By the labeling we adopted, then,  $i < i'$ . Hence,  $d_{j',i',i} = 1$  by Case 6.

Thirdly, suppose that  $d_{i,j',j} = 1$  because of Case 5. If  $m_i \in M_1$ , there exists  $w_{j''}$  such that  $(m_i, w_{j''}) \in c$  and  $d_{i,j',j''} = 1$  because of Case 3, and  $d_{j',i',i} = 1$  by the previous result. If  $m_i \in M_n$  ( $n > 1$ ), then we have shown in Claim 3 that  $(m_{i'}, w_{j'}) \in V$  implies that  $m_{i'} \in M_k$  with  $k > n$ . Hence  $d_{j',i',i} = 1$  because of Case 5.

Finally, note that we cannot have  $d_{i,j',j} = 1$  because of Case 7 since  $(m_i, w_j) \in V$ .  $\square$

Given the orientation  $d$  we have constructed, define two collections of partial orders,  $(\tilde{P}_m : m \in M)$  and  $(\tilde{P}_w : w \in W)$  where we say that  $w\tilde{P}_mw'$  when  $d_{m,w,w'} = 1$  and that  $m\tilde{P}_mw'$  when  $d_{w,m,m'} = 1$ . By Lemma 10, these are well-defined strict partial orders.

Now define the preference of man type  $m$  to be some complete strict extension of  $\tilde{P}_m$  to  $W$ , and similarly for the women. By Lemma 11, these preferences rationalize the matching  $X$ .

The previous construction assumes that  $X$  has one minimal cycle. If  $X$  has more than one minimal cycle, these must not be connected in the graph. Therefore, if we partition the graph into connected components, there will be at most one minimal cycle in each.

In particular, we can partition the set of vertices  $V$  of  $X$  to be  $V = V_1 \cup \dots \cup V_N$  and  $V_m \cap V_n = \emptyset$ . All vertices in each  $V_n$  are connected, but no pair of vertices in different sets are connected. The partition corresponds to the connected components of the graph.

Now re-label the indices of types such that matching  $X$  is a diagonal block matrix:

$$X = \begin{pmatrix} X_1 & O & \cdots & O \\ O & X_2 & \cdots & O \\ \vdots & \vdots & \cdots & \vdots \\ O & O & \cdots & X_N \end{pmatrix}$$

All vertices in  $V_n$  are positive elements in  $X_n$ , and vice versa.

Let  $\tilde{M}_n$  ( $\tilde{W}_n$ ) be the set of types  $m$  ( $w$ ) of men (women) who have a positive elements  $x_{m,w}$  in  $X_n$ . The previous construction, applied to each  $X_n$  separately, yields a rationalizing preference profile of each  $X_n$ , which we denote by  $\tilde{P}$ . For each  $m \in \tilde{M}_n$ , we define a partial

order  $P_m$  on  $W$  to agree with  $\tilde{P}_m$  by adding relations that any  $w \in \tilde{W}_n$  is preferred to every  $w \in W \setminus \tilde{W}_n$ . We similarly define partial orders for the other types of men and types of women. Subsequently, we define  $m$ 's preferences over  $W$  to be a complete extension of  $P_m$ . Women types' preferences are defined analogously.

The resulting profile of preferences rationalizes  $X$  because if  $(v, v')$  is an antiedge with  $v, v' \in V_n$ , for some  $n$ , then (3) is satisfied by the previous construction of preferences, and if  $v$  and  $v'$  are in different components of the partition of  $V$ , then (3) is satisfied because any agent ranks an index in their component over an index in a separate component.

## APPENDIX C. PROOF OF THEOREM 4

**C.1. Proof of necessity.** Let  $X$  be a matching that is rationalizable by the matrix  $\mathcal{A}$ . Suppose for contradiction that the graph  $(V, L)$  associated to  $X$  has a minimal cycle  $c = \langle v_0, \dots, v_N \rangle$ .

We say that an edge  $((m_i, w_j), (m_{i'}, w_{j'})) \in L$  is *vertical* if  $j = j'$  and that it is *horizontal* if  $i = i'$ . Since the cycle  $c$  is minimal, a horizontal edge in  $c$  must be followed by a vertical edge; and a vertical edge in  $c$  must be followed by a horizontal edge (Lemma 3). Thus  $c$  has an even number of vertices. Since  $v_0 = v_N$ , this implies that  $N$  is an even number.

Consider the matching  $X'$ , which coincides with  $X$  on all entries except the ones in  $c$ . For the entries that are vertices in  $c$ , let

$$\begin{aligned} x'_{v_{2n-1}} &= x_{v_{2n-1}} + \varepsilon, & n = 1, \dots, \frac{N}{2} \\ x'_{v_{2n}} &= x_{v_{2n}} - \varepsilon, & n = 0, \dots, \frac{N}{2} - 1 \end{aligned}$$

where  $0 < \varepsilon < \min_{v \in c} x_v$ .

Fix a row  $m_i$  of  $X'$ . For each column  $w_j$ , if  $v_n = (m_i, w_j)$  for some  $n$ , then (modulo  $N$ ) either  $v_{n-1}$  or  $v_{n+1}$  share the same  $w_j$ . Without loss of generality, say that  $v_{n+1}$  shares the same  $w_j$ . By definition of  $X'$ , then  $x_{v_n} + x_{v_{n+1}} = x'_{v_n} + x'_{v_{n+1}}$ . Thus  $\sum_w x'_{i,w} = \sum_w x_{i,w}$ . A similar argument implies that, for each  $w_j$ ,  $\sum_m x'_{m,j} = \sum_m x_{m,j}$ . Hence  $X'$  is a feasible matching in program (2).

Since  $\mathcal{A}$  rationalizes  $X$ , we have that  $\sum_{m,w} \alpha_{m,w} x_{m,w} > \sum_{m,w} \alpha_{m,w} x'_{m,w}$ . Thus,

$$(4) \quad \sum_{m,w} \alpha_{m,w} (x'_{m,w} - x_{m,w}) = \varepsilon \left( \sum_{n=1, \dots, \frac{N}{2}} \alpha_{v_{2n-1}} - \sum_{n=0, \dots, \frac{N}{2}-1} \alpha_{v_{2n}} \right) < 0$$

But then we can consider the matching  $X''$  defined as

$$\begin{aligned} x''_{v_{2n-1}} &= x_{v_{2n-1}} - \varepsilon, & n = 1, \dots, \frac{N}{2} \\ x''_{v_{2n}} &= x_{v_{2n}} + \varepsilon, & n = 0, \dots, \frac{N}{2} - 1, \end{aligned}$$

on the vertices of  $c$ , and which coincides with  $X$  on all entries that are not vertexes of  $c$ .

By the same argument we made for  $X'$ ,  $X''$  is feasible in program (2).

Now, Equation (4) implies that

$$\sum_{m,w} \alpha_{m,w} (x''_{m,w} - x_{m,w}) = \varepsilon \left( - \sum_{n=1, \dots, \frac{N}{2}} \alpha_{v_{2n-1}} + \sum_{n=0, \dots, \frac{N}{2}-1} \alpha_{v_{2n}} \right) > 0;$$

a contradiction of  $X$  being rationalized by  $\mathcal{A}$ .

**C.2. Proof of sufficiency.** Suppose that  $X$  is a matching such that the associated graph contains no cycles. We construct  $\mathcal{A}$  as  $\alpha_{m,w} = \mathbf{1}\{x_{m,w} > 0\}$ , and prove that  $\mathcal{A}$  rationalizes  $X$ .

Clearly,  $\sum_{m,w} \alpha_{m,w} x_{m,w} = \sum_{m,w} x_{m,w}$ . Suppose that  $X'$  is a matching such that  $X'$  is feasible in program (2) for  $X$ , and that  $\sum_{m,w} \alpha_{m,w} x'_{m,w} \geq \sum_{m,w} x_{m,w}$ . We shall prove that  $X' = X$ .

Given  $\mathcal{A}$  as surplus matrix,  $\sum_{m,w} x_{m,w}$  is the maximal surplus that can be achieved in Program (2). To see this, note that all pairs who are matched generate the same value: 1 if they are a pair that is matched under  $x_{m,w}$  and 0 otherwise. The number of different types of men is  $\sum_{m,w} x_{m,w}$  ( $= \sum_m \sum_w x_{m,w}$ ). The number of different types of women is also  $\sum_{m,w} x_{m,w}$  ( $= \sum_w \sum_m x_{m,w}$ ). Thus there are at most  $\sum_{m,w} x_{m,w}$  pairs that can be formed. The maximum value in (2) obtains when all of them generate a surplus of 1. Thus we have  $\sum_{m,w} \alpha_{m,w} x'_{m,w} = \sum_{m,w} x_{m,w}$ .

As a consequence,  $x'_{m,w} = 0$  when  $x_{m,w} = 0$ . Otherwise we would have a pair  $(m, w)$  that are generating a surplus of 0 under  $\mathcal{A}$ , and we cannot have  $\sum_{m,w} \alpha_{m,w} x'_{m,w} = \sum_{m,w} x_{m,w}$ . Thus  $x'_{m,w} = 0$  for all  $(m, w) \notin V$ .

We shall assume that  $(V, L)$  has exactly one connected component. When that assumption fails, we can apply the argument in the sequel to each component separately.

Choose a vertex  $v_0$  in  $V$ . Since  $(V, L)$  contains no cycle, for each  $v \in V$  there is a unique path connecting  $v_0$  to  $v$  in  $(V, L)$ . Let  $\eta(v)$  be the length of the path connecting  $v_0$  to  $v$ . We shall prove the result by induction on  $\eta(v)$ . Specifically, we show that for each  $v$  with maximal  $\eta$ , either the row or the column of  $v$  must be identical in both  $X$  and  $X'$ . We can then consider the submatrix that omitting that row or column, and repeat our argument.

Specifically, define a partial order  $\succ$  on  $V$ , such that  $v_1 \succ v_2$  if and only if  $v_1$  is on the unique path from  $v_0$  to  $v_2$ . Then  $(V, \succ)$  defines a set of maximal chains denoted as  $\{V_1, \dots, V_l\}$ . Each maximal chain has a unique vertex with highest value of  $\eta(v)$ . The following argument can be made for each of these chains.

Let  $(m, w)$  be a vertex with a maximal value of  $\eta(v)$ . Since  $\eta(v)$  is maximal, one of the following two cases hold:

- (1) there is no  $m'$  with  $((m, w), (m', w)) \in L$ , and
- (2) there is no  $w'$  with  $((m, w), (m, w')) \in L$ .

That is, there are either no horizontal edges, or no vertical edges, incident to  $(m, w)$ .

Suppose that Case 1 holds, so  $x_{h,j} = 0$  for all  $h \neq i$ . Then,  $x'_{h,j} = 0$  for all  $h \neq i$ , and  $\sum_h x_{h,j} = \sum_h x'_{h,j}$ , imply that  $x_{m,w} = x'_{m,w}$ . Thus, column  $w_j$  in both matrices  $X'$  and  $X$  coincide.

Consider the submatrices  $X_{\setminus j}$  and  $X'_{\setminus j}$ , obtained after eliminating column  $w_j$ . Then  $\alpha_{\setminus j}$  is the canonical matching of  $X_{\setminus j}$ ; an entry of  $X'_{\setminus j}$  is 0 when the corresponding entry of  $X_{\setminus j}$  is 0, and

$$\sum_{(i,h):h \neq j} \alpha_{i,h} x'_{i,h} = \sum_{(i,h):h \neq j} \alpha_{i,h} x_{i,h}.$$

Finally, the resulting graph  $(V_{\setminus j}, L_{\setminus j})$  contains no cycle.

Similarly, when Case 2 holds, row  $m_i$  of both matrices must coincide. We can then consider the submatrices obtained after eliminating row  $m_i$ .

By applying the above argument to this sequence of submatrices, we show that  $x'_{m,w} = x_{m,w}$  for all  $(m,w) \in V$ . We have already shown that  $x'_{m,w} = x_{m,w} = 0$  for all  $(m,w) \notin V$ . Hence  $X = X'$ .

#### APPENDIX D. PROOF OF THEOREM 3

We proceed by proving first that rationalizability as either  $M$ - or  $W$ -optimal (i.e. extremal-rationalizability) stable matching implies the absence of cycles. Second, we prove that the absence of cycles implies rationalizability as a unique stable matching. Since a unique stable matching is trivially both  $M$ - and  $W$ -optimal, the result follows.

**D.1. Proof that extremal-rationalizability implies the absence of cycles.** Let  $X$  be a matching that is extremal-rationalizable. There exists a preference profile  $P$  such that  $X$  is  $M$ -optimal or  $W$ -optimal stable matching in  $\langle M, W, P, K \rangle$ .

Suppose for contradiction that the graph  $(V, L)$  associated to  $X$  has a minimal cycle  $c = \langle v_0, \dots, v_N \rangle$ . As before, we denote  $m_n$  for the type of the men in  $v_n$ , and  $w_n$  for the type of the women in  $v_n$ , respectively. By Lemma 2 no vertex appears twice in  $c$ , and by Lemma 3 we get the following:

$$\begin{aligned} (v_n, v_{n+1}) \in L \text{ is vertical} &\Rightarrow (v_{n+1}, v_{n+2}) \in L \text{ is horizontal,} \\ (v_n, v_{n+1}) \in L \text{ is horizontal} &\Rightarrow (v_{n+1}, v_{n+2}) \in L \text{ is vertical.} \end{aligned}$$

A preference profile  $(P_{m_i}, P_{w_j})$  defines an orientation  $d$  by setting  $d_{j,i,l} = 1 \iff m_i P_{w_j} m_l$  and  $d_{i,j,k} = 1 \iff w_j P_{m_i} w_k$ . Let  $d$  be the orientation defined by the preference profile  $P$  that rationalizes  $X$  as an extremal matching. According to Lemmas 4 and 5, we can index  $c$  such that the path  $\langle v_n \rangle_{n=0}^{N-1}$  is a flow for  $d$ , and for all  $n = 0, 1, \dots, N-1$  if edge  $(v_n, v_{n+1})$  is vertical (i.e.  $w_n = w_{n+1}$ ), we have  $d_{n,n,n+1} = 0$ , and when the edge is horizontal, we have  $d_{n,n,n+1} = 0$ .

In the following proof, we show that we can make the types of men (women) weakly better (worse) off by “rematching” agents whose matches are involved in the cycle  $c$  while preserving stability. We can also make types of women (men) weakly better (worse) off with a similar rematching. Therefore,  $X$  is neither M-optimal nor W-optimal stable matching.

We capture “rematching,” using a matrix of differences in matches: let  $\mathcal{E}$  be the set of all  $|M| \times |W|$  matrices  $E$  such that for all  $i$  and  $j$ :

- (1)  $e_{i,j} = 0$  if  $(i, j)$  is not in the cycle  $c$ ;
- (2)  $\sum_{h=1}^{|W|} e_{i,h} = 0$ ,  $\sum_{l=1}^{|M|} e_{l,j} = 0$ ; and
- (3)  $\forall (i, j) \quad |e_{i,j}| \leq x_{i,j}$ .

**Claim 4.** *For all  $E \in \mathcal{E}$ , the matrices  $X + E$  and  $X - E$  are stable in  $\langle M, W, P, K \rangle$ ; and either*

- $X - E \leq_M X \leq_M X + E$  and  $X - E \geq_W X \geq_W X + E$ , or
- $X + E \leq_M X \leq_M X - E$  and  $X + E \geq_W X \geq_W X - E$ .

*Proof.* For any  $E \in \mathcal{E}$ ,  $X + E$  is a well defined matching: by Property (2) the row and column sum of  $X + E$  respect the feasibility constraints; by Property (1) and (3), the entries of  $X + E$  are non-negative. The matrix  $X + E$  is also a stable matching, as

$$x_{i,j} + e_{i,j} > 0 \Rightarrow x_{i,j} > 0.$$

Indeed, if there were a blocking pair of type  $m_i$  and type  $w_j$  under  $X + E$ , it would also be a blocking pair under  $X$ . Since  $E \in \mathcal{E} \Rightarrow -E \in \mathcal{E}$ ,  $X - E$  is also well defined and stable.

Observe that as a consequence of Properties (1)-(2),  $e_n$ , which is  $e_{m_n, w_n}$ , alternates in sign. So that if  $e_n \geq 0$  then  $e_{n+1}$ , which is  $e_{m_{n+1}, w_{n+1}}$ , is less than or equal to 0; and  $e_n > 0$  then  $e_{n+1} < 0$ . This implies, first, that if  $e_n = 0$  for some  $n$  then  $E = 0$ . And, second, that one of the following two cases has to hold. (a) For all  $n$ , if  $m_n = m_{n+1} = m$  then  $e_{m, w_n} > 0$  and  $e_{m, w_{n+1}} < 0$ , and if  $w_n = w_{n+1} = w$  then  $e_{m_n, w} < 0$  and  $e_{m_{n+1}, w} > 0$ . (b) For all  $n$ , if  $m_n = m_{n+1} = m$  then  $e_{m, w_n} < 0$  and  $e_{m, w_{n+1}} > 0$ , and if  $w_n = w_{n+1} = w$  then  $e_{m_n, w} > 0$  and  $e_{m_{n+1}, w} < 0$ .

Clearly, if  $E = 0$  then there is nothing to prove. We shall proceed by assuming that we are in case (a), and we shall prove that  $X - E \leq_M X \leq_M X + E$ . It will become clear that if we instead assume that we are in case (b), we would establish that  $X + E \leq_M X \leq_M X - E$ .

Fix  $m \in M$ . By definition of minimal cycle, there is at most one  $n$  such that  $v_n, v_{n+1} \in c$  and  $m_n = m_{n+1} = m$ . If no such  $n$  exists, by Property (1) of  $\mathcal{E}$ ,  $(X - E)_m = (X + E)_m = X_m$ . Thus  $(X - E)_m \leq_m X_m \leq_m (X + E)_m$ .

On the other hand, if there is  $v_n, v_{n+1} \in c$  such that  $m_n = m_{n+1} = m$ , then  $(v_n, v_{n+1})$  is horizontal and  $(v_{n+1}, v_{n+2})$  is vertical. From the orientation  $d$ , we have  $d_{m, n, n+1} = 1$ , implying that  $w_n P_m w_{n+1}$ .

In  $E_m$ , only  $e_n$  and  $e_{n+1}$  are non-zero, and  $0 < e_n = -e_{n+1}$ , as we have assumed that we are in case (a). By definition of  $\leq_m$ ,  $w_n P_m w_{n+1}$  implies that  $(X - E)_m \leq_m X_m \leq_m (X + E)_m$ .

Since the type  $m$  was arbitrary, we obtain  $X - E \leq_M X \leq_M X + E$ . By Theorem 1, this also implies that  $X - E \geq_W X \geq_W X + E$ .

□

**D.2. Proof that the absence of cycles implies unique rationalizability.** We prove that if the graph  $(V, L)$  associated to  $X$  has no cycles, then there is a preference profile  $P$  such that  $(M, W, P, K)$  has  $X$  as its unique stable matching. The matching  $X$  is therefore both  $M$ - and  $W$ -optimal.

We introduce a particular set of preferences. Let  $U = (u_{m,w}) \in \mathbf{R}_+^{|M| \times |W|}$  in which  $u_{m,w} \neq u_{m',w'}$  for all  $(m, w) \neq (m', w')$ . For each  $m$  and  $w$ , a man of type  $m$  and a woman of type  $w$  both receive utility  $u_{m,w}$  by being matched to each other. We denote by  $P_U$  the preference profile induced by such utilities, called a *perfectly correlated preference profile*.

**Lemma 12.** *If a preference profile is perfectly correlated, there exists a unique stable matching.*

*Proof.* Suppose for contradiction that  $X$  and  $Y$  are two distinct stable matchings. Let  $\mathcal{U}$  be the set of numbers  $u_{m,w}$  for  $m$  and  $w$  such that  $x_{m,w} \neq y_{m,w}$ . Let  $(m^*, w^*)$  be such that  $u_{m^*, w^*} \in \mathcal{U}$  and  $u_{m^*, w^*} \geq u$  for all  $u \in \mathcal{U}$ . Suppose, without loss of generality, that  $x_{m^*, w^*} < y_{m^*, w^*}$ . Note that

$$\begin{aligned} \sum_{m: m P_{w^*} m^*} x_{m, w^*} &= \sum_{m: m P_{w^*} m^*} y_{m, w^*} \\ \sum_{w: w P_{m^*} w^*} x_{m^*, w} &= \sum_{w: w P_{m^*} w^*} y_{m^*, w}, \end{aligned}$$

because  $m P_{w^*} m^* \Rightarrow u_{m, w^*} > u_{m^*, w^*} \Rightarrow x_{m, w^*} = y_{m, w^*}$  by construction of  $(m^*, w^*)$ , and similarly for the second equality.

Then,

$$\begin{aligned} \sum_{m: m^* P_{w^*} m} x_{m, w^*} &= K_{w^*} - \sum_{m: m R_{w^*} m^*} x_{m, w^*} \\ &= K_{w^*} - x_{m^*, w^*} - \sum_{m: m P_{w^*} m^*} x_{m, w^*} \\ &> K_{w^*} - y_{m^*, w^*} - \sum_{m: m P_{w^*} m^*} y_{m, w^*} \\ &\geq 0, \end{aligned}$$

as  $x_{m^*, w^*} < y_{m^*, w^*}$ . Similarly,  $\sum_{w: w^* P_{m^*} w} x_{m^*, w} > 0$  and  $(m^*, w^*)$  is a blocking pair of  $X$ ; which contradicts the stability of  $X$ . □

We prove the result by using the absence of cycles to assign cardinal utilities  $U = (u_{m,w})$  so that agents' preferences are perfectly correlated. Then Lemma 12 guarantees that  $X$  is the unique stable matching. We first prove the case when all nodes in  $V$  are connected, and later generalize to the case where there are multiple connected components of  $(V, L)$ .

Suppose that the graph  $(V, L)$  associated to  $X$  has no minimal cycles; so it has no cycles. Choose a vertex  $v_0$  in  $V$ . Since  $(V, L)$  contains no cycles, for each  $v \in V$  there is a unique minimal path connecting  $v_0$  to  $v$  in  $(V, L)$ . Let  $\eta(v)$  be the length of the minimal path connecting  $v_0$  to  $v$ .

We construct correlated preferences that rationalize  $X$  by constructing numbers  $U = (u_{m,w})$ . For  $v \in V$  (i.e.  $x_v > 0$ ),  $u_v = (1 + \eta(v)) + \varepsilon_v$ , and for all other  $(m, w)$  with  $x_{m,w} = 0$ ,  $u_{m,w} = \varepsilon_{m,w}$ . All  $\varepsilon_v$  and  $\varepsilon_{m,w}$  are positive, and distinct real numbers; we assume all  $\varepsilon_{m,w}$  are small enough that if  $\eta(v) > \eta(v')$  then  $u_v > u_{v'}$ . Specifically, let  $(\varepsilon_{m,w})$  be a collection of distinct real numbers such that  $0 < \varepsilon_{m,w} < 1/3$  for all  $(m, w)$ .<sup>14</sup>

Suppose that a type  $m$  man and a type  $w$  woman both receive the same utility  $u_{m,w}$  by being matched to each other. We show that  $X$  is a stable matching in  $\langle M, W, P_U, K \rangle$ . It follows that  $X$  is the unique stable matching because preferences are correlated (Lemma 12).

Suppose for contradiction that a pair  $(m_i, w_j)$  blocks  $X$ . There exist  $m_{i'}$  and  $w_{j'}$  such that  $x_{i,j'} > 0$ ,  $x_{i',j} > 0$ , and  $u_{i,j} > u_{i,j'}$  and  $u_{i,j} > u_{i',j}$ . Since  $x_{i,j'} > 0$  and  $x_{i',j} > 0$ , they are nodes in  $V$ , and  $u_{i,j'} > 1$  and  $u_{i',j} > 1$ . Thus  $u_{i,j} > \max\{u_{i',j}, u_{i,j'}\} > 1$ , by definition of  $U$ , which implies  $x_{i,j} > 0$ . Therefore,  $\langle (m_{i'}, w_j), (m_i, w_j), (m_i, w_{j'}) \rangle$  is a path.

There are unique paths from  $v_0$  to each  $(m_{i'}, w_j)$ ,  $(m_i, w_j)$ , and  $(m_i, w_{j'})$ . Note that  $u_{i,j} > u_{i',j}$  implies that  $\eta((m_i, w_j)) \geq \eta((m_{i'}, w_j))$ . Observe that if  $(v, v') \in L$  then  $\eta(v) \neq \eta(v')$  because if we had  $\eta(v) = \eta(v')$  then  $v$  would not lie in the path  $\langle v_0, \dots, v' \rangle$  and  $v'$  would not lie in the path  $\langle v_0, \dots, v \rangle$ , so  $(v, v') \in L$  would imply the existence of a cycle. So we establish that  $\eta(v) \neq \eta(v')$ . Thus  $\eta((m_i, w_j)) \geq \eta((m_{i'}, w_j))$  implies that

$$(5) \quad \eta((m_i, w_j)) > \eta((m_{i'}, w_j)).$$

Now, (5) is only possible if the unique path from  $v_0$  to  $(m_i, w_j)$  contains  $(m_{i'}, w_j)$ . Then  $\eta(m_i, w_{j'}) > \eta(m_i, w_j)$ ; but this contradicts that  $u_{i,j} > u_{i,j'}$ . Consequently,  $(m_i, w_j)$  cannot be a blocking pair.

When  $(V, L)$  has multiple components,  $\{(V_1, L_1), \dots, (V_N, L_N)\}$ , we can partition  $M$  and  $W$  as  $(M_1, \dots, M_N)$  and  $(W_1, \dots, W_N)$  such that for all  $v \in V_n$ ,  $m_v \in M_n$  and  $w_n \in W_n$ .

For each  $(V_n, L_n)$  with associated sets  $M_n$  and  $W_n$ , we assign utilities  $(u_{m,w})_{(m,n) \in M_n \times W_n}$  similar to the single component case. For other  $m$  and  $w$ , we assign  $u_{m,w} = \varepsilon_{m,w}$ . For all  $(m, w)$ ,  $\varepsilon_{m,w}$  are small and positive real number, and  $\varepsilon_{m,w} \neq \varepsilon_{m',w'}$  when  $(m, w) \neq (m', w')$ .

<sup>14</sup>We use  $\varepsilon_v$  and  $\varepsilon_{m,w}$  only to ensure strict preferences.

Suppose a type  $m$  man and a type  $w$  woman are not matched under  $X$ . If there is  $n$  such that  $(m, w) \in M_n \times W_n$ , then  $(m, w)$  is not a blocking pair by the proof above for the case of a single connected component. If  $(m, w) \in M_n \times W_l$  with  $n < l$ , then, by the construction of  $u_{m,w}$ ,  $w' P_m w$  for any  $w'$  with  $x_{m,w'} > 0$ . Thus  $(m, w)$  is again not a blocking pair;  $X$  is stable matching. By Lemma 12 it is the unique stable matching.

## APPENDIX E. PROOF OF THEOREM 5

Let  $X$  be a rationalizable aggregate matching such that all cycles of the associated graph  $(V, L)$  are balanced. Direct the edges of  $(V, L)$  such that each cycle is oriented as follows: if  $\langle v_0, \dots, v_N \rangle$  is a cycle, then the edge  $(v_n, v_{n+1}) \in L$  is oriented such that  $d(v_{n+1}, v_n) = 1$ , which we denote by  $v_n \rightarrow v_{n+1}$ . For each path  $\langle v_0, \dots, v_N \rangle$ , direct the edges in a similar way. If the matching  $X$  is rationalizable, then such an orientation of the edges exists and defines a rationalizing preferences profile  $P$  (the sufficiency proof of Theorem 2). The rationalizing preferences have the property that if  $x_{i,j} = 0$  and  $x_{i',j'} > 0$  then  $w_{j'} P_{m_i} w_j$  if  $i = i'$ , and  $m_{i'} P_{w_j} m_i$  if  $j = j'$ .

First, if  $X$  has no cycles, then it is rationalizable as the unique stable matching (Theorem 3), so there is nothing to prove, as a unique stable matching is also the median stable matching. Suppose then that  $X$  has at least one cycle  $c = \langle v_0, \dots, v_N \rangle$ . Enumerate the vertexes of the cycle such that  $v_n \rightarrow v_{n+1}$  in the orientation (directed graph) of  $(V, L)$  above, and  $v_0$  lies in the same row as  $v_1$ . Let

$$\Theta = \min \{v_0, v_2, \dots, v_{N-2}\} = \min \{v_1, v_3, \dots, v_{N-1}\}.$$

Let  $\mathcal{E}$  be the set of all  $|M| \times |W|$  matrices  $E$  of integer numbers such that

- $e_{i,j} = 0$  if  $(i, j)$  is not in the cycle  $c$ ;
- for all  $i$  and  $j$ ,  $\sum_h e_{i,h} = 0$   $\sum_l e_{l,j} = 0$ ; and
- $|e_{i,j}| \leq \Theta$ .

We want to make two observations about the matrices in  $\mathcal{E}$ . First,  $(X + E)_{i,j} > 0 \Rightarrow x_{i,j} > 0$ , so  $X + E$  is a stable matching for all  $E \in \mathcal{E}$ . Second,  $E \in \mathcal{E}$  if and only if  $-E \in \mathcal{E}$ ; and  $E \in \mathcal{E}$  is such that  $X_i \leq_{m_i} (X + E)_i$  if and only if  $(X - E)_i \leq_{m_i} X_i$ . Similarly,  $E \in \mathcal{E}$  is such that  $X_j \leq_{w_j} (X + E)_j$  if and only if  $(X - E)_j \leq_{w_j} X_j$ .

We need to prove that there are no other stable matchings than the ones obtained through matrices in  $\mathcal{E}$ : Then  $X$  is a median stable matching.

Let  $Y \neq X$  be another stable matching in the resulting market  $\langle M, W, P, K \rangle$ . We shall prove that  $y_{i,j} \neq x_{i,j}$  only if  $x_{i,j}$  is a vertex in a minimal cycle of  $X$ . Suppose then that  $y_{i,j} \neq x_{i,j}$ . The number of single agents of each type is the same in  $X$  as in  $Y$  (Proposition 8; in this case it is zero, as  $X$  has no single agents). So, if  $x_{i,j} < y_{i,j}$  then there is  $h \neq j$  and

$l \neq i$  such that  $y_{i,h} < x_{i,h}$  and  $y_{l,j} < x_{l,j}$ . Similarly, if  $x_{i,j} > y_{i,j}$  then there is  $h \neq j$  and  $l \neq i$  such that  $y_{i,h} < x_{i,h}$  and  $y_{l,j} < x_{l,j}$ .

We can apply the previous observation repeatedly to obtain a sequence  $(i_1, j_1), \dots, (i_N, j_N)$  with  $(i_1, j_1) = (i_N, j_N)$  such that for each  $n \pmod N$ :

- (1)  $(x_{i_n, j_n} - y_{i_n, j_n})(x_{i_{n+1}, j_{n+1}} - y_{i_{n+1}, j_{n+1}}) < 0$
- (2)  $i_n \neq i_{n+1} \iff j_n = j_{n+1}$ .

**Claim 5.** For  $n = 1, \dots, N$ ,  $x_{i_n, j_n} > 0$ .

Suppose for contradiction that  $0 = x_{i_1, j_1} < y_{i_1, j_1}$ . Without loss of generality, assume that  $i_1 = i_2$ . By definition of the rationalization  $P$ , we have that  $w_{j_2} P_{m_{i_1}} w_{j_1}$ , as  $x_{i_1, j_1} = 0$  and  $x_{i_1, j_2} > y_{i_1, j_2} \geq 0$ . We can now show that if  $(i_n, j_n)$  and  $(i_{n+1}, j_{n+1})$  differ in  $i$ , then  $j_n$  prefers  $m_{i_{n+1}}$  to  $m_{i_n}$ ; and that if they differ in  $j$ , then  $i_n$  prefers  $w_{j_{n+1}}$  to  $w_{j_n}$ . This fact, which we prove in the next paragraph, establishes the contradiction:  $i_N \neq i_{N-1}$ , but  $m_{i_{N-1}} P_{w_{i_N}} m_{i_N}$  by definition of  $P$  and because  $0 = x_{i_N, j_N} = x_{i_1, j_1}$ .

To prove the fact, we reason by induction. We have already established that  $w_{j_2} P_{m_{i_1}} w_{j_1}$ . Suppose that  $w_{j_n} P_{m_{i_n}} w_{j_{n-1}}$ . By Property 1 of the sequence  $\langle (i_n, j_n) \rangle_{n=1}^N$ , either  $x_{i_{n-1}, j_{n-1}} > 0$  and  $x_{i_{n+1}, j_{n+1}} > 0$ ; or  $y_{i_{n-1}, j_{n-1}} > 0$  and  $y_{i_{n+1}, j_{n+1}} > 0$  (or both hold). Then the stability of  $X$  and  $Y$  implies that  $m_{i_{n+1}} P_{w_{j_n}} m_{i_n}$ . The proof for the case when  $w_{i_n} P_{w_{i_n}} m_{i_{n-1}}$  is similar.

The claim implies that the sequence  $\langle v_n \rangle = \langle x_{i_n, j_n} \rangle$  is a cycle in  $(V, L)$ . Thus a stable  $Y$  can only differ from  $X$  in vertexes that are part of a cycle of  $(V, L)$ . Let  $E = Y - X$ ; we shall prove that  $E \in \mathcal{E}$ . We established above that  $e_{i,j} \neq 0$  only if  $x_{i,j}$  is a vertex in a cycle. We now prove that  $|e_{i,j}| \leq \Theta$ . Clearly,  $e_{i,j} \geq -x_{i,j} \geq -\Theta$ . We show that if  $e_{i,j} > 0$  then there is  $h$  such that  $e_{i,j} + e_{i,h} = 0$ .

If  $e_{i,j} > -e_{i,h}$  for all  $h \neq i$  then there is  $h_1$  and  $h_2$  such that some men of type  $m_i$  who are married to women of type  $h_1$  and  $h_2$  in  $X$  are married to women of type  $w_j$  in  $Y$ . Then we can define two cycles, and  $x_{i,j}$  would be a vertex in both of them. The first cycle has  $(x_{i,j}, x_{i,h_1})$  as the first edge, and the remaining edges defined inductively, by the definition of  $\langle (i_n, j_n) \rangle$  above. The second cycle has  $(x_{i,j}, x_{i,h_2})$  as the first edge, and the remaining edges defined inductively. The resulting two cycles would be connected, which contradicts the hypothesis that  $X$  is rationalizable. So there must exist some  $h$  with  $e_{i,j} \leq -e_{i,h}$ . An analogous argument applied to  $e_{i,h}$  implies that  $e_{i,j} \geq -e_{i,h}$ ; so  $e_{i,j} = -e_{i,h}$ . Then,  $e_{i,j} \leq \Theta$ , as  $e_{i,h} \geq -\Theta$ .

## APPENDIX F. EXAMPLES

**F.1. Example for the sufficiency proof of rationalizability.** The following example is rationalizable using many different preference profiles. The algorithm used in the proof of Theorem 2 can only construct some of them.

Consider the following aggregate matching.

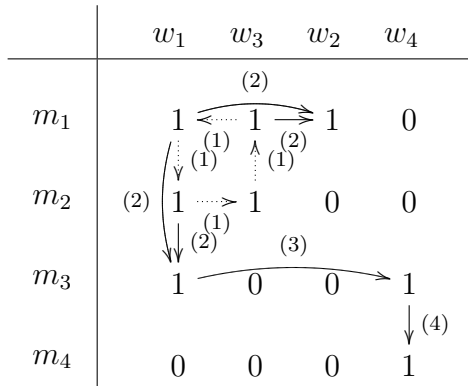
$$X = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

We illustrate the algorithm used in the proof of Theorem 2.

There is a minimal cycle,  $\{(m_1, w_1), (m_4, w_1), (m_4, w_3), (m_1, w_3)\}$ . From the cycle, we obtain  $\bar{M}_1 = \{m_1, m_4\}$  and  $\bar{W}_1 = \{w_1, w_3\}$ . Subsequently following the proof, we define

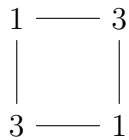
- $\bar{M}_2 = \{m_2\}$ ,  $\bar{W}_2 = \{w_2\}$ ,
- $\bar{M}_3 = \emptyset$ ,  $\bar{W}_3 = \{w_4\}$ ,
- and  $\bar{M}_4 = \{m_3\}$ ,  $\bar{W}_4 = \emptyset$ .

Following the steps in the proof, we obtain a graph summarizing preferences.



All orientations labeled (1) are determined by the minimal cycle. The orientations denoted (2), (3), and (4) are sequentially determined as we apply the algorithm.

**F.2. There are matchings rationalizable, but not median-rationalizable.** A rationalizable aggregate matching that cannot be rationalized as a median stable matching:



As we show in Remark 1, the only way to rationalize this aggregate matching is to have a cycled preference profile, i.e., if  $w_1 P_{m_1} w_2$  then  $m_2 P_{w_1} m_1$ ,  $w_2 P_{m_2} w_1$ , and  $m_1 P_{w_2} m_2$ . Similarly, if  $w_2 P_{m_1} w_1$  then the preferences of agents are cycled in the opposite direction. Regardless of the choice, all the feasible aggregate matchings are stable, so there are 5 stable aggregate matchings in total. Therefore, the median aggregate stable matching is the one where each

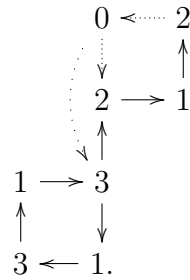
agent is matched to both possible partners twice, which is

$$\begin{array}{ccc} 2 & \text{---} & 2 \\ | & & | \\ 2 & \text{---} & 2 \end{array}$$

**F.3. Median-rationalizability condition is not necessary.** The following example shows that the sufficient condition in Theorem 5 is not necessary. Suppose that there are four types of men, three types of women, and consider an aggregate matching

$$X = \begin{pmatrix} 0 & 0 & 2 \\ 0 & 2 & 1 \\ 1 & 3 & 0 \\ 3 & 1 & 0 \end{pmatrix}.$$

We choose the preferences such that the graph corresponding to  $X$  is as follows. Note that we indicate preference with an arrow, so that for example  $x_{i,j} \rightarrow x_{i,h}$  means that  $w_h P_{m_i} w_j$ .



Note that there is a cycle  $\{(m_3, w_1), (m_3, w_2), (m_4, w_2), (m_4, w_1)\}$  and that this cycle is not balanced.

By “rotating” the cycle  $\{(m_3, w_1), (m_3, w_2), (m_4, w_2), (m_4, w_1)\}$  in a clockwise direction we obtain the aggregate matching

$$\begin{pmatrix} 0 & 0 & 2 \\ 0 & 2 & 1 \\ 0 & 4 & 0 \\ 4 & 0 & 0 \end{pmatrix},$$

which is better for the men. However, by counterclockwise rotations we obtain the matchings

$$\begin{pmatrix} 0 & 0 & 2 \\ 0 & 2 & 1 \\ 2 & 2 & 0 \\ 2 & 2 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 2 \\ 0 & 2 & 1 \\ 3 & 1 & 0 \\ 1 & 3 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 2 \\ 0 & 2 & 1 \\ 4 & 0 & 0 \\ 0 & 4 & 0 \end{pmatrix},$$

which are all better for women than  $X$ .

Now, *with the rationalization in the arrows above*, the following “joint” rotation of the cycle and the upper entries of the matchings is stable as well:

$$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 4 & 0 \\ 4 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 3 \\ 0 & 4 & 0 \\ 4 & 0 & 0 \end{pmatrix}.$$

These two matchings are better for the men. In all, then, under the rationalizing preferences as in the arrows, there are 7 stable matchings, and  $X$  is the median stable matching.

There is a crucial aspect of the example that makes this possible. Note that, if we are to rotate the upper part of the graph, we need preferences to be as indicated by the arrows. In particular, we must have  $x_{2,3} \rightarrow x_{1,3}$  for the graph to be rationalizable; then, to accommodate a  $> 0$  entry in  $x_{1,2}$  after the rotation, we must have  $x_{1,3} \rightarrow x_{1,2}$ , or otherwise we would get a blocking pair  $(m_1, w_3)$ . However, positive entries in  $x_{1,3}$  and in  $x_{3,2}$  imply that we need  $x_{1,2} \rightarrow x_{3,2}$  (the long dotted arrow in the graph in order to satisfy that  $x_{1,3} \rightarrow x_{1,2}$ . Now, a modification of  $X$  that has a positive entry in  $x_{1,2}$  is only possible if we simultaneously set  $x_{3,1} = 0$ , as  $x_{3,1} \rightarrow x_{3,2}$  and  $x_{1,2} \rightarrow x_{3,2}$ . Hence the rotation of the upper side of the graph *is not feasible under any of the modification of  $X$  that improve the matches of the women*.

There is, therefore, an asymmetry in the graph that allows us to offset the unbalancedness of the number of men and women in the cycle.

## REFERENCES

- ABDULKADIROĞLU, A., P. PATHAK, A. ROTH, AND T. SÖNMEZ (2005): “The Boston Public School Match,” *American Economic Review*, 95(2), 368–371.
- ABDULKADIROĞLU, A., P. A. PATHAK, AND A. E. ROTH (2005): “The New York City High School Match,” *The American Economic Review*, 95(2), pp. 364–367.
- ADACHI, H. (2000): “On a Characterization of Stable Matchings,” *Economic Letters*, 68, 43–49.
- ALKAN, A., AND D. GALE (2003): “Stable schedule matching under revealed preference,” *Journal of Economic Theory*, 112(2), 289 – 306.
- BEARMAN, P., J. MOODY, AND K. STOVEL (2004): “Chains of affection: The structure of adolescent romantic and sexual networks,” *The American Journal of Sociology*, 110(1), 44–91.
- BECKER, G. S. (1973): “A Theory of Marriage: Part I,” *Journal of Political Economy*, 81(4), 813–846.
- BLUNDELL, R., M. BROWNING, AND I. CRAWFORD (2003): “Nonparametric Engel Curves and Revealed Preference,” *Econometrica*, 71(1), 205–240.

- CHAMBERS, C. P., AND F. ECHENIQUE (2009): “The core matchings of markets with transfers,” Working Papers 1298, California Institute of Technology, Division of the Humanities and Social Sciences.
- CHOO, E., AND A. SIOW (2006): “Who Marries Whom and Why,” *Journal of Political Economy*, 114(1), 175–201.
- CURRARINI, S., M. JACKSON, AND P. PIN (2009): “An economic model of friendship: Homophily, minorities, and segregation,” *Econometrica*, 77(4), 1003–1045.
- DAGSVIK, J. K. (2000): “Aggregation in Matching Markets,” *International Economic Review*, 41(1), 27–57.
- ECHENIQUE, F. (2008): “What Matchings Can Be Stable? The Testable Implications of Matching Theory,” *Mathematics of Operations Research*, 33(3), 757–768.
- ECHENIQUE, F., S. LEE, AND M. SHUM (2010): “Aggregate Matchings,” Caltech SS Working Paper 1328.
- (2011): “The Money Pump as a Measure of Revealed Preference Violations,” Caltech SSWP 1328.
- ECHENIQUE, F., S. LEE, AND M. B. YENMEZ (2010): “Existence and Testable Implications of Extreme Stable Matchings,” Caltech SS Working Paper 1337.
- ECHENIQUE, F., AND J. OVIEDO (2004): “Core Many-to-one Matchings by Fixed Point Methods,” *Journal of Economic Theory*, 115(2), 358–376.
- (2006): “A Theory of Stability in Many-to-Many Matching Markets,” *Theoretical Economics*, 1(2), 233–273.
- ECHENIQUE, F., AND L. YARIV (2010): “An Experimental Study of Decentralized Matching,” mimeo Caltech.
- ECHENIQUE, F., AND M. B. YENMEZ (2007): “A Solution to Matching with Preferences over Colleagues,” *Games and Economic Behavior*, 59(1), 46–71.
- FLEINER, T. (2003): “A Fixed-Point Approach to Stable Matchings and Some Applications,” *Mathematics of Operations Research*, 28(1), 103–126.
- FOX, J. T. (2007): “Estimating Matching Games with Transfers,” mimeo: University of Chicago.
- GALE, D., AND L. S. SHAPLEY (1962): “College Admissions and the Stability of Marriage,” *The American Mathematical Monthly*, 69(1), 9–15.
- GALICHON, A., AND B. SALANIE (2009): “Matching with Trade-offs: Revealed Preferences over Competing Characteristics,” mimeo., Ecole Polytechnique.
- HATFIELD, J., AND P. MILGROM (2005): “Auctions, Matching and the Law of Auctions Matching and the Law of Aggregate Demand,” *American Economic Review*, 95(4), 913–935.

- HYLLAND, A., AND R. ZECKHAUSER (1979): “The efficient allocation of individuals to positions,” *The Journal of Political Economy*, 87(2), 293–314.
- KELSO, A. S., AND V. P. CRAWFORD (1982): “Job Matching, Coalition Formation, and Gross Substitutes,” *Econometrica*, 50, 1483–1504.
- KESTEN, O., AND U. ÜNVER (2009): “A Theory of School Choice Lotteries,” mimeo, Boston College and Carnegie Mellon University.
- KLAUS, B., AND F. KLIJN (2006): “Median stable matching for college admissions,” *International Journal of Game Theory*, 34(1), 1–11.
- (2010): “Smith and Rawls share a room: stability and medians,” *Social Choice and Welfare*, 35, 647–667.
- KOMORNIK, V., Z. KOMORNIK, AND C. VIAUROUX (2010): “Stable schedule matchings by a fixed point method,” *UMBC Economics Department Working Papers*.
- KÜÇÜKŞENEL, S. (2011): “Core of the Assignment Game via Fixed Point Methods,” *Journal of Mathematical Economics*, 47(1), 72–76.
- OSTROVSKY, M. (2008): “Stability in Supply Chain Networks,” *American Economic Review*, 98(3), 897–923.
- ROTH, A., U. ROTHBLUM, AND J. VANDE VATE (1993): “Stable matchings, optimal assignments, and linear programming,” *Mathematics of Operations Research*, 18(4), 803–828.
- ROTH, A., AND M. SOTOMAYOR (1988): “Interior Points in the Core of Two-Sided Matching Markets,” *Journal of Economic Theory*, 45, 85–101.
- (1990): *Two-sided Matching: A Study in Game-Theoretic Modelling and Analysis*, vol. 18 of *Econometric Society Monographs*. Cambridge University Press, Cambridge England.
- ROTH, A. E. (2008): “Deferred Acceptance Algorithms: History, Theory, Practice and Open Questions,” *International Journal of Game Theory*, 36(3), 537–569.
- ROTHBLUM, U. (1992): “Characterization of stable matchings as extreme points of a polytope,” *Mathematical Programming*, 54(1), 57–67.
- SCHWARZ, M., AND M. B. YENMEZ (2011): “Median Stable Matching for Markets with Wages,” *Journal of Economic Theory*, 146(2), 619–637.
- SHAPLEY, L., AND M. SHUBIK (1971): “The assignment game I: The core,” *International Journal of Game Theory*, 1(1), 111–130.
- SÖNMEZ, T., AND U. ÜNVER (Forthcoming): “Matching, Allocation, and Exchange of Discrete Resources,” in *Handbook of Social Economics*, ed. by J. Benhabib, A. Bisin, and M. Jackson. Elsevier.
- TEO, C.-P., AND J. SETHURAMAN (1998): “The geometry of fractional stable matchings and its applications,” *Mathematics of Operations Research*, 23(4), 874–891.

- VANDE VATE, J. H. (1989): “Linear programming brings marital bliss,” *Operations Research Letters*, 8(3), 147–153.
- VARIAN, H. R. (1985): “Non-Parametric Analysis of Optimizing Behavior with Measurement Error,” *Journal of Econometrics*, 30, 445–458.