Submission Number: PET11-11-00269

What price stability? Social welfare in matching markets.

James Boudreau
University of Texas - Pan American

Vicki Knoblauch
University of Connecticut

Abstract

In two-sided matching markets, stability can be costly. We define social welfare functions for matching markets and use them to formulate a definition of the price of stability. We then show that it is common to find a price tag attached to stability, and that the price of stability can be substantial. Therefore, when choosing a matching mechanism, a social planner would be well advised to weigh the price of stability against the value of stability, which varies from market to market.

Submitted: March 15, 2011.

What Price Stability? Social Welfare in Matching Markets*

James W. Boudreau†
Dept. of Economics and Finance
University of Texas-Pan American

and

Vicki Knoblauch‡
Dept. of Economics
University of Connecticut

April, 2011

Abstract

In two-sided matching markets, stability can be costly. We define social welfare functions for matching markets and use them to formulate a definition of the price of stability. We then show that it is common to find a price tag attached to stability, and that the price of stability can be substantial. Therefore, when choosing a matching mechanism, a social planner would be well advised to weigh the price of stability against the value of stability, which varies from market to market.

JEL Classification Codes: C78, D63
*Keywords:*Price of stability; social welfare functions; matching.

*We are grateful to Bettina Klaus for helpful comments and suggestions. We would also like to thank seminar participants at Rice University and the UECE Lisbon Meetings 2010: Game Theory and Applications, at which a related paper was presented.

†Email: boudreaujw@utpa.edu.
‡Email: vicki.knoblauch@uconn.edu.

1 Introduction

We treat formally two issues that have been dealt with in an *ad hoc* manner in the marriage matching literature: the price of stability and welfare comparisons of matchings.

In many applications stability is vital. By definition, every unstable matching contains at least one blocking pair, a man and a woman who prefer each other to their assigned mates. A blocking pair has the potential to cause a matching to unravel. However, in some scenarios stability is not vital; for example, a strong central authority can prevent renegotiation, that is, can prevent blocking pairs from abandoning their assigned mates to form new marriages, and can thereby prevent unstable matchings from unraveling. In particular, a school district with a strong central administration can forbid schools from altering their enrollments once matchings have been made by the mechanism of choice. Alternatively, if participants are made aware that they would expect to do better under a mechanism that doesn't guarantee stability than under any stable matching, they might be willing and able to commit as a group to non-renegotiation. Even when participants are not aware that they could expect to do better under a mechanism that doesn't guarantee stability, they may be extremely unhappy about the results of a stable matching mechanism.[1] Even when forced or committed non-renegotiation cannot be attained, it may be so difficult for blocking pairs to find each other that unraveling is unlikely to start and once started is likely to stall. In cases such as these where stability is not vital, it is useful to weigh the price of stability against the value of stability. Finally, even when stability is vital, curiosity alone motivates an investigation of the price of stability. With all these motives in hand, we will begin our investigation by adapting the definition of price of stability from Roughgarden and Tardos (2007) for use in the marriage matching arena.

This brings us to our second issue. Before we can present a formal definition of the price of stability, we need a general method for assigning values to matchings. Many

[1] For example, in 2002 a class-action lawsuit was brought against the National Residency Matching Program, the matching mechanism that assigns new doctors to residencies, on the grounds that the stable mechanism held down the salaries of new doctors. Bulow and Levin (2006) provide theoretical support for the defendants' claim, and Crawford (2008) suggests improvements to the matching procedure to alleviate such problems. More details on the case itself are available in those papers.

studies assign values to matchings, but this has always been done in an *ad hoc* manner. For example, Ergin (2002) and Klaus and Klijn (2007) observe that stability and Pareto optimization are inconsistent when we define Pareto dominance with consideration for only one side of the market, in other words, in markets for which we are interested in pleasing the agents on only one side of the market. From the fact that the authors make this observation, we can infer that they place, albeit implicitly, a higher value on stable matchings than on unstable matchings and a higher value on Pareto optimal matchings than on Pareto dominated matchings. (Incidentally, their observation is also an example of an informal consideration of the price of stability.) Similarly, the observation by Roth (1982) that stability and strategy-proofness are incompatible in certain circumstances is also an example of informal assignment of values to matchings and informal consideration of the price of stability. Many other instances are discussed in our literature review in the next section.

We will provide a formal foundation for assigning values to matchings by adapting the concept of a social welfare function for use in the marriage matching arena. Probably the chief argument against using social welfare functions to assign values to matchings is that men's and women's preferences are ordinal in nature, and therefore any comparison between matchings that takes men's and women's preferences into account should not be cardinal in nature. It will be convenient to delay the presentation of several counterarguments until after we have defined social welfare functions and given several examples.

We can now give a brief informal preview of our two key definitions. A social welfare function (SWF) assigns a nonnegative real number to every ordered pair consisting of a matching and a preference profile. For a fixed market size, the price of stability associated with a SWF f is the maximum over all preference profiles of the ratio of the maximum value of f over all matchings and the maximum value of f over all stable matchings. Then we will say there is a price tag attached to stability if the price of stability as just defined is greater than one. Our definitions differ from the models on which they were based only in that we allow the value of a SWF to vary not only with the outcome of some process (in our case the outcome is a matching), but also with a feature of the market, the preference profile. This makes it possible for the values we place on matchings to depend, in a variety

of ways, on the levels of satisfaction of the participants.

The rest of the paper is organized as follows. Section 2 reviews the literature. In Section 3 we define social welfare functions (SWFs) over marriage matchings, provide several examples and argue that SWFs are legitimate tools for the study of marriage matching. In Section 4 we define the price of stability for marriage matching and show that for most of our examples of SWFs there is a price tag attached to stability. We show that for at least two of our examples a search for that price tag does not require the construction of strange, rare or complex preference profiles. Rather, for large markets a randomly chosen preference profile will almost certainly show that the SWF in question comes with a price tag attached to stability. We demonstrate that the price of stability for two of our examples is substantial. Section 5 concludes with a discussion of how simulation can provide information about the price of stability for many scenarios that are intractable via theory. Proofs appear in the appendix.

2 Literature Review

Although our study is the first to make a formal, systematic study of social welfare and of the price of stability in two-sided matching markets, many authors have discussed the relative merits of different matching mechanisms and in doing so have of necessity dealt with social welfare and at times with the price of stability, albeit in an *ad hoc* manner. Unfortunately, different papers often use similar or identical terms ("fairness", for example) in reference to notions of welfare that are actually quite different from one another. Thus, in this section we review the literature to compare and contrast these different notions of welfare, allowing us to put existing results, as well as the results of this paper, in proper context.

The most common notion of equilibrium in matching markets is that of *stability*. In a traditional marriage matching market, a match is stable, if the match has no *blocking pairs*. A blocking pair is a man and women who mutually prefer one another to their existing partners.

Stability is often considered to be the top priority in matching markets. Many successive

rounds of recontracting can be costly, so quickly arriving at a stable match can save time and effort. Also, since stable outcomes are Pareto-efficient, stability is often thought of as synonymous with general *efficiency* in matching markets. In almost all of the matching literature, then, stability is assumed to be critical for market welfare, although it is not always explicitly designated as such.

Gale and Shapley (1962) proved that all marriage markets have at least one stable outcome, and provided a (now eponymous) algorithm to obtain such an outcome. Subsequent work by Kelso and Crawford (1982) and Roth (1984) proved that stable outcomes are also always possible in many-to-one and many-to-many matching markets respectively, given fairly weak separability restrictions on agents' preferences. Echenique and Oviedo (2006) then considered several related variants of stability for many-to-many markets, and verified slightly stronger preference restrictions which again ensure the existence of such outcomes.

In all instances, however, though a matching market's set of stable outcomes can be guaranteed non-empty, it is rarely single-valued. And when more than one stable matching does exist for a market, the set has a lattice structure with respect to the interests of the two sides of the market (for the case of one-to-one matching see Roth and Sotomayor, 1990, Theorem 3.8 attributed to Conway; for generalizations to the many-partner cases see Blair, 1988, and Echenique and Oviedo, 2006). That is, in the case of marriage matching, there will always be one stable matching that is most preferred by all men and least preferred by all women. This is known as the *man-optimal* matching. There will also be an analogous *woman-optimal* matching, with matchings in between being partially ordered by at-least-as-good-for-every-man when moving from the man-optimal toward the woman-optimal matching and vice versa.

This polarization of interest in stable matchings points to another potential dimension of welfare for matchings in addition to stability, one sometimes referred to as "fairness". For example, if the aforementioned Gale-Shapley algorithm is used in practice it constructs either the man- or woman-optimal matching, but showing such favoritism to one side of the market may seem unfair to the unfavored side. Thus, there has been a substantial amount of interest in matchings that attain some compromise between the interests of the two sides. If stability is taken as a necessary criteria that must be satisfied, this gives rise

to the notion of *median stable matchings*, which have been studied by Teo and Sethuraman (1998), Sethuraman et al. 2006), and Klaus and Klijn (2006a, 2010).

Loosely speaking, a median stable matching is a matching that is stable, but which favors neither side of the market over the other (to the extent that is possible). More specifically, if there are $2k + 1$ stable matchings for a two-sided market, each agent will have a (possibly weak) preference ordering over them. The median stable matching is such that it is every agent's $(k + 1)$th most preferred out of the set of all stable matchings. If there is an even number of stable matchings, say $2k$, the median gives all agents their kth or $(k + 1)$th choice out of the set of stable matchings. Teo and Sethuraman (1998) and Sethuraman et al. (2006) prove the existence of such matchings for one-to-one and many-to-one settings respectively using linear programming methods. Klaus and Klijn (2010) and (2006a) do the same, though using simpler methods based on the lattice structure of the set of stable matchings.

Romero-Medina (2005) similarly studies compromise between the two sides of the market, defining what he calls an *equitable* algorithm, which selects only those matchings that are in the middle of the lattice of stable matchings. That is, those matchings that are in between the polarized interests of the two sides of the market. Romero-Medina (2005) labels this set of matchings the *equitable set*, but the concept obviously bears a close resemblance to the idea of median matchings. In all five cases (Teo and Sethuraman, 1998; Sethuraman et al., 2006; Klaus and Klijn, 2010, 2006a; Romero-Medina, 2005), the concept of the compromise between the two sides of the market is specifically labeled as a type of fairness.

In particular, a unique median stable matching (or equitable matching) represents a type of *endstate fairness*, since all agents end up with their $(k+1)$th-most preferred stable partner out of the set of $2k + 1$ stable matchings. Clearly this specific type of fairness may not be possible, however, unless the number of stable matchings is odd. Furthermore, in the case of randomized stable matching mechanisms–literally mechanisms that select randomly from among the set of all stable matchings–evaluating an end state may not be meaningful. If preferences are strictly ordinal, an induced probability distribution over the set of stable matchings (a randomized mechanism's end state) is difficult to interpret, in

terms of fairness or otherwise. Thus, in the absence of endstate fairness, Klaus and Klijn (2006b) explore what they refer to as *procedural fairness*.

Procedurally fair randomized stable mechanisms are those in which each agent's placement in the procedure is uniformly random. This in turn means that each agent has an equal chance at selecting their most-preferred of all stable partners, an equal chance at selecting their second-most-preferred stable partner, and so on. Thus, although the end state may not be considered fair *ex post*, the procedure is fair *ex ante*. Klaus and Klijn (2006b) compare two different procedurally fair randomized mechanisms, as well as a third which is actually a randomized extension of Romero-Medina's (2005) equitable algorithm, and which combines elements of both procedural and end-state fairness.

Whether considering endstate or procedural fairness, Klaus and Klijn (2006a, 2006b, 2010) and Romero-Medina (2005) frequently cite philosopher John Rawls' (1971) conception of justice. This is with good reason, since equality of opportunity is an integral part of Rawls' conception, and his brand of justice has proved to have a considerable and longstanding impact on many strands of literature. Another aspect to Rawlsian justice, however, and one that is perhaps more often and more closely associated with Rawls in the economics literature, is what is sometimes known as the "difference principle" or the *maximin* criterion.

The maximin criterion of social welfare simply seeks to make the worst-off individual as well off as possible. Although such a criterion is not without drawbacks (cf. Dasgupta, 1974, or Tungodden, 1999), its presence has nevertheless persisted, perhaps due to the appealing logic of Rawls (1971) himself. Paraphrasing very loosely, Rawls (1971) envisions an initial position of ignorance in which an individual is without any knowledge of where they will end up in the hierarchy of society. If such an indvidual is given the ability to choose society's division of assets from that position, a perfectly equal society might seem most appealing, but to the extent that any inequality is present it should only be in the interest of making the worst-off individual(s) as well off as possible.

Unsurprisingly, this maximin criterion has made its way into the matching literature. Masarani and Gokturk (1989) investigate what they refer to as "fair" matchings, but require four specific criteria to qualify as fair. A fair matching mechanism according to

Masarani and Gokturk (1989) must satisfy what they label (*i.*) *gender indifference* (neither side of the market should be favored over the other, a type of endstate fairness, in line with the aforementioned equitability or median matchings), (*ii.*) *peer indifference* (the order of individuals in the process should not matter, in line with the aforementioned procedural fairness), (*iii.*) *maximin optimality* (the aforementioned Rawlsian criterion), and (*iv.*) *stability*. Unfortunately, this definition of fairness proves impossible to satisfy with an unrestricted domain of preferences (Masarani and Gokturk, 1989).

Masarani and Gokturk correctly identify a conflict between the maximin criterion and stability in the midst of their other requirements (a result we will confirm here in isolation from the other two qualifications), and so they then relax their definition of fairness to include only elements *i.*, *ii.*, and *iv*, giving stability primacy over the Rawlsian maximin principle. Still, even with this more restricted definition, fairness proves to be impossible to guarantee for all matching markets with an unrestricted domain of preferences. Again, we will confirm that even in isolation, gender indifference and stability are impossible to guarantee together for all matching markets.

Though it may be arguable in the case of the maximin criterion, all of the welfare criteria that we have reviewed so far have been ordinal in nature, relying only on agents preferring one match to another. Alternatively, welfare could also be measured by interpreting some kind of cardinality in individuals' preferences, a practice that also has precedence in the literature.

The most straightforward utilitarian welfare measure for matching markets is a simple summation of the rankings individuals give their partners. An individual matched with their first-ranked partner counts as one, an individual matched with their second-ranked partner counts as two, and so on. A matching with a lower *choice count* is then considered preferable to one with a higher choice count. Combining stability with utilitarianism, it is then possible to choose a *minimum choice count* stable matching (McVitie and Wilson, 1971), also known as the *egalitarian* stable matching (Gusfield and Irving, 1989). But in fact, the minimum choice count over all stable matchings will not always be the minimum choice count out of all matchings, and in this paper we will show how far apart the two can be.

The utilitarian concept of the choice count can also be combined with notions of fairness or justice, as well as stability. Romero-Medina (2001), for example, defines the concept of *envy* for individuals in matching markets as the number of other agents matched with partners the individual would prefer to their own. In other words, the amount of envy an individual has in a match is simply the ranking they give to their partner in that match minus one. An individual matched with their third ranked partner therefore envies two other people.

Romero-Medina (2001) then defines the *sex equal* matching as the stable matching that minimizes the difference between the cumulative amount envy on each side of the market. In a classic marriage market, this is equivalent to minimizing (within the set of stable matchings) the difference between the total men's and women's choice count. The sex equal stable matching is therefore the cardinal analog to the ordinal concepts of median matchings or Romero-Medina's (2005) equitable matchings, aiming at compromise between the two sides of the market.

Of course, Romero-Medina (2001) only selects from the set of stable matchings, claiming that stability is a requirement for "fairness". We will show explicitly here that, as it does with most criteria, the requirement of stability conflicts with the goal of minimizing envy across the two sides of the market, a property we call *balancedness*. More generally, Klaus (2009) defines a *fair* matching as one that is both stable and envy-free (meaning no agent is any better or worse off than another), and proves that such matchings are impossible to guarantee in marriage matching markets, even with monetary transfers allowed.

Finally, there are also cases in which only the welfare of one side of the market is of concern. Balinski and Sönmez (1999), Ergin (2002) and Klaus and Klijn (2007) study student-school matching problems in which the students' welfare is the only concern. Schools are treated as objects with fixed "priorities" rather than agents with manipulable preferences. They define efficiency as Pareto-efficiency from the sole perspective of the student side of the market, a form of stability for the student side of the market, and fairness as the traditional notion of stability, taking into account the preferences of students and the priorities of schools. Ergin (2002) and Klaus and Klijn (2007) then establish conditions on preferences to ensure the existence of what they call fair and efficient matching mechanisms.

The common theme in all of these previous works is the primacy of stability over all other criteria. Our major departure in this paper is the fact that we consider the sacrifices, in terms of other welfare criteria, that must be made if stability is to be maintained. We consider how great the cost of stability can potentially be, and how likely it is that such a cost will have to be incurred given a random set of preferences.

Our goal in this paper is not to suggest that stability is an undesirable property. It's merits are clear. Rather, our goal is simply to point out the fact that it does entail tradeoffs, and to begin to clarify what those tradeoffs are.

3 Social Welfare Functions

The model considered here is the simple marriage matching problem first popularized by Gale and Shapley (1962). The model features two finite disjoint sets of agents denoted $M = \{m_1, m_2, \ldots, m_n\}$ and $W = \{w_1, w_2, \ldots, w_n\}$. We adopt the marriage market interpretation and refer to the two sets as men and women, but alternative interpretations categorize agents as firms and workers, or workers and machines. Each agent has a complete, strict, transitive preference ordering over the agents on the other side of the market. Man i's preferences are given by a one to one and onto ranking function r_{m_i}: $W \to \{1, 2 \ldots, n\}$ where w_j is preferred to w_k by m_i if $r_{m_i}(w_j) < r_{m_i}(w_k)$. Woman j's preferences are similarly represented by r_{w_j}. A market's ranking profile, P_n, is then simply the collection of all agents' preference orderings induced by their ranking functions.

The outcome of a marriage matching problem is a matching of men and women given by a one-to-one and onto function μ: $M \to W$. A matching is said to be *stable* if there does not exist a *blocking pair* $\{m_i, w_j\}$ such that $r_{m_i}(w_j) < r_{m_i}(\mu(m_i))$ and $r_{w_j}(m_i) < r_{w_j}(\mu(w_j))$. As proved by Gale and Shapley (1962), at least one such matching always exists. Let S denote the set of all stable matchings. Let μ_M and μ_W be the men-propose and the women-propose Gale-Shapely matching, respectively.

Definition 1. A *social welfare function (SWF) f assigns a positive real number to every ordered triple* (n, μ_n, P_n)*, where* P_n *and* μ_n *are, respectively, a ranking profile and a matching for* $M \cup W$ *with* $|M| = |W| = n$.

We will write $f(\mu_n, P_n)$ rather than $f(n, \mu_n, P_n)$ and we will sometimes write μ rather than μ_n and/or P rather than P_n if n is fixed or if it is not necessary to specify n.

We now introduce three examples of SWFs and nine examples of families of SWFs. The first two examples are formalizations of the priority placed on stability and Pareto efficiency which, as we mentioned in our literature review, is implicit in much of the existing work on matching.

Example 1.

$$f(\mu, P_n) = \begin{cases} 2 & \text{if } \mu \text{ is stable} \\ 1 & \text{otherwise} \end{cases}$$

Example 2.

$$f(\mu, P_n) = \begin{cases} 2 & \text{if } \mu \text{ is Pareto efficient} \\ 1 & \text{otherwise} \end{cases}$$

Example 3. the utilitarian SWF

$$f(\mu, P_n) = \frac{1}{\sum_{a \in M \cup W} r_a(\mu(a))}$$

As we also mentioned in our literature review, the utilitarian SWF appears in the marriage matching literature as well; its denominator is the sum of the number of proposals in the men-propose and the number of proposals in the women-propose Gale-Shapley algorithm (the choice count). Also, its denominator is the notion of envy described by Romero-Medina (2001). It is a simple measure of aggregate satisfaction.

Families of Social Welfare Functions.

Our first few examples of families of SWFs reflect the possibility that some priority may be given to specific individuals or groups of individuals (entire genders), culminating in the general notion of monotonicity.

Example 4. f is a-focused for some $a \in M \cup W$; that is, $f(\mu, P_n) > f(\mu', P_n)$ if $r_a(\mu(a)) < r_a(\mu'(a))$

Example 5. f is strictly female monotonic; that is, $f(\mu, P_n) > f(\mu', P_n)$ if μ female Pareto

dominates μ', that is, if no female prefers μ' to μ and some female prefers μ to μ'.

Example 6. f is strictly male monotonic.

Example 7. f is strictly monotonic; that is, $f(\mu, P_n) > f(\mu', P_n)$ if μ Pareto dominates μ', that is, if no male or female prefers μ' to μ and some male or female prefers μ to μ'.

Our next family of SWFs, Rawlsian SWFs, relate to the maximin optimality of Masarani and Gokturk (1989) as follows: for P fixed, μ is maximin optimal if and only if μ maximizes a Rawlsian SWF.

Example 8. f is Rawlsian; that is, $f(\mu, P_n) > f(\mu', P_n)$ if
$\max\{r_a(\mu(a))\colon\ a \in M \cup W\} < \max\{r_a(\mu'(a))\colon\ a \in M \cup W\}$.

Our next example combines the spirit of democracy with the context of matching. Given P_n, we define a *dominating set* of matchings as one such that each member defeats every matching outside the set in a pairwise election. We then define the *Smith set* as the smallest nonempty dominating set; that is, the smallest nonempty set of matchings such that each member defeats every matching outside the set in a head to head election where the electorate is $M \cup W$.

The Smith set exists for every P_n since the set of all matchings is a dominating set, the dominating sets are nested, and the number of matchings is finite.

Example 9. f is Smith; that is, for every P, f is maximal on the Smith Set.

Our next example is simply a generalization of the already-mentioned priority of stability to a family of SWFs.

Example 10. f respects stability; that is, $f(\mu, P_n) > f(\mu', P_n)$ if μ is stable and μ' is not (for example, Example 1).

Finally, our last two examples allude to the notions of envy-freeness or gender-equality mentioned in the literature review.
Example 11. f respects *gender balancedness*; that is, $f(\mu, P_n) > f(\mu', P_n)$ if
$\left|\sum_{w \in W} r_w(\mu(w)) - \sum_{m \in M} r_m(\mu(m))\right| < \left|\sum_{w \in W} r_w(\mu'(w)) - \sum_{m \in M} r_m(\mu'(m))\right|$.

The concept of gender balancedness is in the spirit of Romero-Medina (2001), seeking to minimize the envy across the two sides of the market. Alternatively, it could instead be the priority to maintain a balance across all individuals.

Example 12. f respects *balancedness* across individuals; that is, $f(\mu, P_n) > f(\mu', P_n)$ if $\sum_{a\in M\cup W}\sum_{b\in M\cup W/a}|r_a(\mu(a)) - r_b(\mu(b))| < \sum_{a\in M\cup W}\sum_{b\in M\cup W/a}|r_a(\mu'(a)) - r_b(\mu'(b))|$.

With definition and examples in hand, we are ready to argue that it is appropriate, that is, reasonable and useful, to consider SWFs whose values may depend on participants' ordinal preferences. Here are three arguments.

1. Ranking profiles are ordinal and don't express intensity of preference, but that doesn't prevent a central authority, social planner or mechanism designer from possessing preferences over matchings that *are* held with varying degrees of intensity.

2. Ranking profiles don't express intensities of preference, but a participant (man or woman) probably does hold preferences over mates or even matchings with various degrees of intensity. These preferences might be expressible by a SWF which could perhaps be used by the participant to decide whether to participate in the market. For example, the preferences of participant a might be represented by an a-focused SWF, while those of a more community minded participant b might be represented by some combination of a b-focused SWF and a Rawlsian SWF.

3. SWFs are powerful tools that generate valid *ordinal* conclusions. For example, a study of the utilitarian SWF often leads to conclusions like "If preferences are similar to what we have seen in the past, the average man can expect a match under μ that he ranks seven places better than his match under μ'." (Boudreau and Knoblauch, 2010)

4 The Price of Stability

For non-cooperative games, the price of stability is defined as the ratio of a game's best equilibrium outcome value to its best (possibly non-equilibrium) outcome value, as mea-

sured by some objective function (Roughgarden and Tardos, 2007, p. 446). We adapt that notion here for a marriage matching environment in light of the concept of SWFs which we have already described.

Definition 2. For each positive integer n, the price of stability for a social welfare function is

$$P\text{of}Sn(f) = \max_{P_n}\Big(\frac{\max\{f(\mu, P_n)\colon\ \mu \text{ is a matching }\}}{\max\{f(\mu, P_n)\colon\ \mu \in S\}}\Big)$$

For a SWF f, we will say there is a price tag attached to stability if $P\text{of}Sn(f) > 1$.

Our investigation of the price of stability will proceed in three stages in which we present three types of evidence to show that it is common for stability to come with a price tag attached and that the price of stability can be substantial. First we prove that $P\text{of}Sn(f) > 1$ for five of our examples. All proofs are consigned to an appendix.

Theorem 1. *There is a price tag attached to stability for Examples 3, 4, 5, 6 8, 11 and 12, $n \geq 3$.*

Remark. $P\text{of}Sn(f) = 1$ trivially for f from Example 1, 10 or 2 (this last since every stable matching is Pareto efficient). One nontrivial example for which $P\text{of}Sn(f) = 1$ is Example 9, for which $P\text{of}Sn(f) = 1$ because it can be shown that the set of all stable matchings is a subset of the Smith set.

Proposition 1. *If a SWF f is Smith, then $PofSn(f) = 1$ for $n \geq 1$.*

Concerning Example 7, $P\text{of}Sn(f) > 1$ holds for some but not all strictly monotonic $SWFs$. For instance, we have already seen that $P\text{of}Sn(f) > 1$ for the utilitarian SWF, but $P\text{of}Sn(f) = 1$ for the sum of the utilitarian SWF and the SWF of Example 1.

Notice that our definition of the price of stability is worst-case based in that we maximize over all preference profiles. Therefore for a given SWF the price of stability might be greater than one but the ratio in the definition might be one for all but a few preference

profiles. In the second stage of our analysis we show that for two of our examples the price tag attached to stability is not due to rare and unusual preference profiles. In fact, for those two examples, for large n, the ratio in the definition of the price of stability is greater than one for nearly all preference profiles, where the meaning of "nearly all" is made precise in the statement of Theorem 2.

Theorem 2. *If f is a-focused and for each n a ranking profile is chosen uniform randomly, or if f is utilitarian and for each n men's preferences are identical and women's are chosen uniform-randomly, then*

$Prob(max\{f(\mu, P_n)\colon\ \mu \text{ is a matching } \} > max\{f(\mu^s, P_n)\colon\ \mu^s \text{is stable}\}) \to 1$ *as* $n \to +\infty$.

In our third and final theorem, we prove that the price of stability can be substantial.

Theorem 3. *If* $n \geq 3$,

for the utilitarian SWF, $P\text{of}Sn(f) \geq n/3$

for the a-focused SWF, $f(\mu, P_n) = 1/r_a(\mu(a))$, $P\text{of}Sn(f) \geq n$

for the Rawlsian SWF, $f(\mu, P_n) = 1/\max\{r_a(\mu(a))\colon a \in M \cup W\}$, $P\text{of}Sn(f) \geq n/2$

5 Concluding Remarks

Theorem 3 demonstrates that the price of stability can be substantial and Theorem 2 showed that a price tag attached to stability need not owe it existence to rare and unusual preference profiles. In real-life markets, a social planner or mechanism designer might want a combination of Theorem 2- and Theorem 3-type information; that is, she might want to know the price of stability defined as an average over a category of preference profiles rather than a maximum over all preference profiles. For example, the social planner might know the approximate level of correlation (the extent to which the preferences are similar) on each side of the market, and/or the approximate level of intercorrelation (the extent to which men prefer women who prefer them) and wish to know the expected-magnitude

price of stability.

A future study will use another tool to address such questions–simulation. The simulation approach involves three steps: the generation of preference profiles, the construction of matchings by the mechanisms to be tested and the evaluation of the social welfare function for each matching. The process is repeated many times to yield an approximation of the expected level of social welfare provided by each mechanism. The social planner then has all the information needed (including the price of stability) to choose a matching mechanism.

The advantage of simulation is that a single approach works for any category of preference profiles and any matching mechanism. On the other hand the theoretical approach used in this study is more rigorous and more transparent.

Appendix

Proof of Theorem 1. We treat the examples one by one.

Example 3. f the utilitarian social welfare function. Define P_n by

$m_1 : w_2, w_1, w_3, ...$
$m_2 : w_2, w_3, w_1, ...$
$m_3 : w_3, w_2, w_1, ...$
$m_k : w_k, ...$ for $k > 3$

$w_1 : m_1, m_2, m_3, ...$
$w_2 : m_1, m_2, m_3, ...$
$w_3 : m_2, m_3, w_1, ...$
$w_k : m_k, ...$ for $k > 3$

Where the ellipses indicate the remaining preferences are irrelevant. Since μ_M is men optimal, μ_W is women optimal and for this P_n $\mu_M = \mu_W$, the unique stable matching is $\mu^s = (m_1, w_2), (m_2, w_3), (m_3, w_1), (m_k, w_k)$ for $k > 3$. Then $f(\mu^s, P_n) = \frac{1}{2n+5}$. For $\mu = (m_k, w_k)$ for all k, $f(\mu, P_n) = \frac{1}{2n+3}$. Therefore $PofSn(f) \geq \frac{2n+5}{2n+3} > 1$.

Example 4. f is a-focused. Without loss of generality, suppose f is m_3 focused. For P_n, μ, and μ^s as in the previous proof, $r_{m_3}(\mu(m_3)) = 1 < 3 = r_{m_3}(\mu^s(m_3))$. Therefore, $f(\mu, P_n) > f(\mu^s, P_n)$ so that P of $S_n(f) > 1$.

Example 8. Suppose f is Rawlsian. For P_n, μ and μ^s as in the previous proofs, $\max\{r_a(\mu(a))\colon\ a \in M \cup W\} = 2 < 3 = \max\{r_a(\mu^s(a))\colon\ a \in M \cup W\}$. Therefore, $f(\mu, P_n) > f(\mu^s, P_n)$ so that $P\text{of}Sn(f) > 1$.

Example 6. f is strictly male monotonic. Define P_n by

$m_1 : w_2, w_1, w_3, ...$
$m_2 : w_1, w_2, w_3, ...$
$m_3 : w_2, w_1, w_3, ...$
$m_k : w_k, ...$ for $k > 3$

$w_1 : m_1, m_3, m_2, ...$
$w_2 : m_2, m_1, m_3, ...$
$w_3 : m_3, ...$
$w_k : m_k, ...$ for $k > 3$

Then the unique stable matching is $\mu^s = \mu^M = \mu^W = (m_k, w_k)$ for $k \geq 1$. Consider $\mu = (m_1, w_2), (m_2, w_1), (m_k, w_k), (m_k, w_k)$ for $k > 2$. Then μ male Pareto dominates μ^s. Therefore $f(\mu, P_n) > f(\mu^s, P_n)$ so that $P\text{of}Sn(f) > 1$.

Example 5. f is strictly female monotonic. Follows from the previous proof by symmetry.

Example 11. f respects *gender balancedness*. Define P_n by

$m_1 : w_1, w_{n-1}, w_n, ...$
$m_2 : w_2, w_n, w_{n-1}, ...$
$m_3 : w_3, w_1, w_{n-2}, ...$
$m_k : w_k, w_{k-2}, w_{n-k+1}, ...$ for $k > 3$

$w_1 : m_1, m_{n-1}, m_n, ...$
$w_2 : m_1, m_2, m_{n-1}, ...$
$w_3 : m_1, m_3, m_{n-2}, ...$

$w_n : m_1, m_k, m_{n-k+1}, ...$ for $k > 3$

The unique stable matching is $\mu^s = \mu^M = \mu^W = (m_k, w_k)$ for $k \geq 1$. In that matching, $r_m(\mu^s(m)) = 1$ for all $m \in M$, $r_{w_1}(\mu^s(w_1)) = 1$, and $r_{w_k}(\mu^s(w_k)) = 2$ for all $k \geq 2$. But the matching $\mu = (m_1, w_n), (m_2, w_{n-1}), (m_3, w_{n-2}), (m_k, w_{n-k+1})$ for $k > 3$ yields $r_a(\mu(a)) = 3$ for all $a \in M \cup W$, making it more balanced both across genders and across individuals.

Example 12. f respects *balancedness*. See previous treatment for example 11. ∎

Proof of Proposition 1. Fix P_n. Suppose $\mu^s \in S$ and μ is a matching. Then for $m \in M$, $r_m(\mu(M)) < r_m(\mu^s(m))$ implies $r_{\mu(m)}(m) > r_{\mu(m)}(\mu^s(\mu(m)))$ or else $(m, \mu(m))$ would be a blocking pair for μ^s, contradicting its stability. It follows that if $\mu^s \in S$ then no matching defeats μ^s in a head to head election. Therefore, S is a subset of the Smith set. Since f is Smith, S is a subset of $\arg\max\{f(\mu, P_n) : \mu \text{is a matching}\}$. Therefore, $PofSn(f) = 1$. ∎

For the proof of theorem 2 we will need a brief description of the men-propose McVitie-Wilson (1971) algorithm, which for any ordering $m_{i_1}, m_{i_2}, \ldots, m_{i_n}$ of men provides an n-round sequence of proposals leading to μ^M. First m_{i_1} proposes to a woman and they form a temporary pair. Before round k, m_{i_1}, m_{i_2}, and $m_{i_{k-1}}$ are engaged. Round k begins when m_{i_k} proposes to his favorite woman. If she is engaged she either rejects m_{i_k} or accepts him and rejects her current partner. The rejected man proposes to his next favorite, and so on. Round k ends when a woman receives her first proposal. At the end of round n, the matching μ^M has been achieved.

Proof of Theorem 2a. Without loss of generality, suppose $a = m_1$ and $r_{m_1}(w_1) = 1$. It is sufficient to prove $\lim_{n \to +\infty} Prob(\mu^M(m_1) \neq w_1) = 1$.

Given $\epsilon > 0$ choose positive integer $K > \frac{1}{\epsilon}$.

Consider three procedures. Procedure 1 is the McVitie-Wilson algorithm with proposing sequence $m_1, m_2, \ldots, m_n$. Procedure 2 is exactly $[(n \ln n)/2]$ proposals long and consists of Procedure 1 truncated after $[(n \ln n)/2]$ proposals if necessary; or finished with m_1 repeatedly proposing to w_1 after w_1 receives her K^{th} proposal or if Procedure 1 ends in fewer

than $[(n \ln n)/2]$ proposals. Procedure 3 is $[(n \ln n)/2] - (K-1)n$ draws with replacement from an urn containing n balls. For $i = 1, 2$ let Q_i be the probability that Procedure i ends with fewer than K proposals to w_1. Let Q_3 be the probability that Procedure 3 ends with fewer than K draws of ball 1.

Then $Q_2 < Q_3$, since first of all in Procedure 2 more than $[(n \ln n)/2] - (K-1)n$ proposals are made with w_1 in play. This holds because w_1 is out of play only if a man is making a proposal after proposing to w_1 and he is not m_1 in the repeated-proposal-to-w_1 mode. There are at most $K-1$ men who make such proposals and each must make fewer than n such proposals. Second, in each proposal with w_1 in play, the probability that w_1 is proposed to is at least $1/n$.

Next, each of he following inequalities holds for sufficiently large n.

$$Q_3 < \sum_{j=1}^{K} \binom{[(n \ln n)/2] - (K-1)n}{j} \left(\frac{1}{n}\right)^j \binom{n-1}{n}^{[(n \ln n)/2]-(K-1)n-j}$$

$$\leq K \binom{[(n \ln n)/2] - (K-1)n}{K} \left(\frac{1}{n}\right)^K \left(\frac{n-1}{n}\right)^{[(n \ln n)/2]-(K-1)n-K}$$

$$\leq K \binom{n \ln n}{K} \left(\frac{1}{n}\right)^K \left(\frac{n-1}{n}\right)^{(n \ln n)/3}$$

$$\leq K \frac{(n \ln n)^K}{K!} \left(\frac{1}{n}\right)^K \left(\frac{1}{\sqrt[4]{e}}\right)^{\ln n}$$

$$(\ln n)^K \frac{1}{\sqrt[4]{n}} < \epsilon$$

Finally, by the definitions of Procedures 1 and 2, if Procedure 1 ends with fewer than K proposals to w_1 and Procedure 2 ends with at least K proposals to w_1, it must be that Procedure 1 ends after fewer than $[(n \ln n)/2]$ proposals. By a result of Pittel (1989, Theorem 2) this occurs with probability less than ϵ for sufficiently large n.

In summary, given $\epsilon > 0$, for sufficiently large n $Q_1 \leq Q_2 + \epsilon \leq Q_3 + \epsilon \leq 2\epsilon$. Since $Prob(\mu^M(m_1) \neq w_1)$ is greater than or equal to the probability that w_1 receives at least $1/\epsilon$ proposals times the probability that $\mu^M(m_1) \neq w_1$ given that w_1 receives at least $1/\epsilon$

proposals, for sufficiently large n, $Prob(\mu^M(m_1) \neq w_1) \geq (1-2\epsilon)(1-\epsilon)$. ∎

Proof of Theorem 2b. Since men's preference are identical an exchange of partners has no effect on men's aggregate satisfaction, so we need to consider only women's aggregate satisfaction.Without loss of generality assume $r_{m_i}(w_n) = n$ for all i.

For P_n as in the hypotheses, the probability that w_n receives her k^{th} choice under the unique stable matching μ^s in $1/n$ for $k = 1, 2, \ldots, n$. Therefore, the probability that w_n is matched with her $[n/\ln n]^{th}$ choice or worse (where $[n/\ln n]$ is the integer part of $n/\ln n$) is

$$1 - \frac{[n/\ln n] - 1}{n} \geq 1 - \frac{n/\ln n}{n} = 1 - \frac{1}{\ln n} \to 1 \text{ as } n \to +\infty$$

Now suppose $\mu^s(w_n)$ is w_n's $[n/\ln n]^{th}$ choice or worse. There are at least $[n/\ln n] - 1$ women matched with men w_n prefers to $\mu^s(w_n)$. We can assume $\mu^s(w_i) = m_i$ and $r_{w_n}(m_i) = i$ for $i = 1, 2, \ldots, [n/\ln n] - 1$ and $\mu^s(w_n) = m_n$. Then for $i = 1, 2, \ldots, [n/\ln n] - 1$ if w_n exchanges partners with w_i, $r_{w_n}(m_n) - r_{w_n}(m_i) \geq [n/\ln n] - i$. The trade results in an increase in aggregate satisfaction if $r_{w_i}(m_n) - r_{w_i}(m_i) < [n/\ln n] - i$, which will happen if $r_{w_i}(m_n) < r_{w_i}(m_i) + [n/\ln n] - i$. Since $\mu^s = \mu^M$ under the hypotheses on preferences and since $r_{m_n}(w_n) = n$, m_n must have been rejected by w_i before proposing to w_n. Therefore $r_{w_i}(m_i) < r_{w_i}(m_n)$. The trade with w_i results in an increase in aggregate satisfaction if $r_{w_i}(m_i) < r_{w_i}(m_n) < r_{w_i}(m_i) + [n/\ln n] - i$. The probability of this occurring is $Prob(r_{w_i}(m_i) \leq r_{w_i}(m_n) < r_{w_i}(m_i) + [n/\ln n] - i) \geq \frac{[n/\ln n]-(i+1)}{n}$, where the inequality follows from the fact that w_i's preferences are uniform random and the fact that if $r_{w_i}(m_i) + [n/\ln n] - i > n$ then the probability in question is 1.

Therefore the probability that none of the trades between w_n and w_i increase net aggregate satisfaction for $i = 1, 2, \ldots, [n/\ln n] - 1$ is at most

$$\Pi_{i=1}^{[n/\ln n]-1}(1 - \frac{n/\ln n - (i+1)}{n}) \leq (1 - \frac{n/2\ln n}{n})^{n/4\ln n}$$

where the inequality follows by considering only slightly more than the first quarter of the factors, each of which is bounded above by $(1 - \frac{n/2\ln n}{n})$. Continuing

$$(1 - \frac{n/2\ln n}{n})^{n/4\ln n} = ((1 - \frac{1}{2\ln n})^{2\ln n})^{n/8\ln^2 n} \to 0 \text{ as } n \to +\infty \text{ since } (1 - \frac{1}{2\ln n})^{2\ln n} \to e$$

as $n \to +\infty$

Finally,

$Pr(max\{f(\mu, P_n)\colon\ \mu \text{ is a matching } \} > max\{f(\mu^s, P_n)\colon\ \mu^s \in S\}) \geq$

$Pr(r_{w_n}(\mu^s(w_n)) \geq [n/\ln n]) \times$

$$Pr(\text{some } w_n - w_i \text{ trade increases } f \mid (r_{w_n}(\mu^s(w_n)) \geq [n/\ln n])$$

$\to 1$ as $n \to +\infty$ since we showed both factors approach 1. ∎

Proof of Theorem 3. Consider the preference profile P_n:

$m_1 : w_2, w_1, ...$
$m_2 : w_2, \ldots, w_3, w_1$
.
.
$m_k : w_k, \ldots, w_{k+1}, w_1$
.
.
$m_{n-1} : w_{n-1}, \ldots, w_n, w_1$
$m_n : w_n, \ldots, w_1$

$w_1 : m_1 \ldots m_n$
$w_2 : m_1, m_2, ...$
.
.
$w_k : m_{k-1}, m_k, ...$
.
.
$w_n : m_{n-1}, m_n, ...$

where the ellipses indicate arbitrary preferences. Since $\mu^M = \mu^W$, there is a unique stable matching μ^s: $(m_1, w_2), (m_2, w_3), \ldots, (m_{n-1}, w_n), (m_n, w_1)$ and for the utilitarian SWF, $f(\mu^s, P_n) = \frac{1}{n^2+2}$. For μ: $(m_1, w_1), (m_2, w_2), \ldots, (m_n, w_n)$, $f(\mu, P_n) = \frac{1}{3n}$. Therefore

$P\text{of}Sn(f) \geq \frac{n^2+2}{3n} \geq \frac{n}{3}$.

For the a-focused SWF $f(\mu, P_n) = \frac{1}{r_a(\mu(a))}$, without loss of generality we take $a = m_n$. Then for P_n, μ and μ^s from the first part of this proof, $f(\mu, P_n) = 1$ and $f(\mu^s, P_n) = \frac{1}{n}$ so that $P\text{of}Sn(f) \geq n$.

For the Rawlsian SWF $f(\mu, P_n) = 1/\max\{r_a(\mu(a))\colon\ a \in M \cup W\}$ for P_n, μ, μ^s as above, $f(\mu^s, P_n) = 1/r_{m_n}(\mu^s(m_n)) = \frac{1}{n}$ and $f(\mu, P_n) = 1/r_{w_n}(\mu(w_n)) = \frac{1}{2}$ so that $P\text{of}Sn(f) \geq n/2$. ∎

References

Balinski, M. and T. Sönmez (1999). A tale of two mechanisms: Student placement. *Journal of Economic Theory 84*, 73–94.

Blair, C. (1988). The lattice structure of the set of stable matchings with multiple partners. *Mathematics of Operations Research 13*(4), 619–628.

Boudreau, J. W. and V. Knoblauch (2010). Marriage matching and intercorrelation of preferences. *Journal of Public Economic Theory 12*(3), 587–602.

Bulow, J. and J. Levin (2006). Matching and price competition. *American Economic Review 96*(4), 652–668.

Crawford, V. P. (2008). The flexible salary match: A proposal to increase the salary flexibility of the national resident matching program. *Journal of Economic Behavior and Organization 66*, 149–160.

Dasgupta, P. (1974). On some problems arising from professor rawls' conception of distributive justice. *Theory and Decision 4*, 325–344.

Echenique, F. and J. Oviedo (2006). A theory of stability in many-to-many matching markets. *Theoretical Economics 1*, 233–273.

Ergin, H. (2002). Efficient resource allocation. *Econometrica 70*(6), 2489–2497.

Gale, D. and L. S. Shapley (1962). College admissions and the stability of marriage. *The American Mathematical Monthly 69*(1), 9–15.

Gusfield, D. and R. W. Irving (1989). *The stable marriage problem, structure and algorithms*. The M.I.T. Press, Cambridge, MA.

Kelso, A. S. and V. P. Crawford (1982). Job matching, coalition formation, and gross substitutes. *Econometrica 50*(4), 1483–1504.

Kesten, O. (2010). School choice with consent. *Quarterly Journal of Economics 125*(3), 1297–1348.

Klaus, B. (2009). Fair marriages: An impossibility. *Economics Letters 105*, 74–75.

Klaus, B. and F. Klijn (2006a). Median stable matching for college admissions. *International Journal of Game Theory 34*, 1–11.

Klaus, B. and F. Klijn (2006b). Procedurally fair and stable matching. *Economic Theory 27*, 431–447.

Klaus, B. and F. Klijn (2007). Fair and efficient student placement with couples. *International Journal of Game Theory 36*, 177–207.

Klaus, B. and F. Klijn (2010). Smith and rawls share a room: stability and medians. *Social Choice and Welfare 35*, 647–667.

Masarani, F. and S. S. Gokturk (1989). On the existence of fair matching algorithms. *Theory and Decision 26*, 305–322.

McVitie, D. G. and L. B. Wilson (1971). The stable marriage problem. *Communications of the ACM 14*(7), 486–490.

Pittel, B. (1989). The average number of stable matchings. *SIAM Journal of Discrete Mathematics 2*, 530–549.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press, Cambridge, MA.

Romero-Medina, A. (2001). Sex-equal stable matchings. *Theory and Decision 50*, 197–212.

Romero-Medina, A. (2005). Equitable selection in bilateral matching markets. *Theory and Decision 58*, 305–324.

Roth, A. E. (1982). The economics of matching: Stability and incentives. *Mathematics of Operations Research 7*(4), 617–628.

Roth, A. E. (1984). Stability and polarization of interests in job matching. *Econometrica 52*(1), 47–57.

Roth, A. E. and M. Sotomayor (1990). *Two-sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge University Press, Cambridge, MA.

Roughgarden, T. and Éva Tardos (2007). Introduction to the inefficiency of equilibria. In N. Nisan, T. Roughgarden, Éva Tardos, and V. V. Vazirani (Eds.), *Algorithmic Game Theory*, pp. 443–459. Cambridge University Press.

Sethuraman, J., C.-P. Teo, and L. Qian (2006). Many-to-one stable matching: Geometry and fairness. *Mathematics of Operations Research 31*(3), 581–596.

Teo, C.-P. and J. Sethuraman (1998). The geometry of fractional stable matchings and its applications. *Mathematics of Operations Research 23*(4), 874–891.