Mathematical Social Sciences 67 (2014) 27–33

Contents lists available at ScienceDirect

Mathematical Social Sciences

journal homepage: www.elsevier.com/locate/econbase

What price stability? Social welfare in matching markets[☆]

James W. Boudreau[a], Vicki Knoblauch[b,*]

[a] *Department of Economics and Finance, University of Texas-Pan American, United States*
[b] *Department of Economics, University of Connecticut, United States*

HIGHLIGHTS

- We formalize the concept of social welfare functions for matching markets.
- Stability can involve a price in terms of other notions of social welfare.
- We show some price tags are very very likely.
- We also show that price tags can be substantial.

ARTICLE INFO

Article history:
Received 22 May 2012
Received in revised form
15 May 2013
Accepted 30 September 2013
Available online 18 October 2013

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.

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 the theory and practice of marriage matching, stability has traditionally been considered a vital requirement. By definition, every unstable matching contains at least one blocking pair, a man and a woman who prefer each other to their assigned mates. The presence of such pairs can be problematic since individuals then have the incentive to leave their assigned partners to be with each other.

But 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 falling apart after partners are assigned. A school district, for example, with a strong central administration, could potentially forbid schools from altering their enrollments once matchings have been made by the mechanism of choice.

In fact, it is important to recognize that two-sided matching markets were originally centralized not to prevent matchings from falling apart *after* assignments have been made, but rather to prevent backward unraveling, that is, to prevent individuals from trying to form partnerships earlier and earlier, before adequate time is allowed to ascertain partners' post-match potential. The requirement of stability, however, is neither necessary nor sufficient to prevent backward unraveling. Unstable matchings can survive over time and successfully limit backward unraveling (Unver, 2001, 2005), while stable mechanisms on their own do not guarantee that backward unraveling will not occur (Halaburda, 2010).

If participants are made aware that they would expect to do better under a mechanism that does not guarantee stability than under any stable matching, they might be willing and able to commit as a group to not seek alternative partners before or after the unstable matching is assigned. Alternatively, even when participants are not aware that they could expect to do better under a mechanism that does not guarantee stability, they may be extremely unhappy about the results of a stable matching mechanism.[1] And

[☆] 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.

* Corresponding author. Tel.: +1 860 486 9076.
E-mail addresses: boudreaujw@utpa.edu (J.W. Boudreau), vicki.knoblauch@uconn.edu (V. Knoblauch).

[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

0165-4896/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.mathsocsci.2013.09.004

even when forced or committed non-renegotiation cannot be attained, it may be so difficult for blocking pairs to find each other that match breaking 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 considered necessary, curiosity alone motivates an investigation of the price of stability.

This paper is not the first to question the priority of stability over other welfare criteria in matching markets. Axtell and Kimbrough (2008), for example, compare the man-optimal stable matching mechanism introduced by Gale and Shapley (1962) with a decentralized algorithm that terminates in unstable matchings and show that the unstable mechanism outperforms significantly in terms of agents' average rank of partner. Anshelevich et al. (2009, 2013) similarly point out that stability can come at a substantial cost in utilitarian terms, and investigate the potential size of such costs both theoretically and with simulation experiments for various distributions of agents' utility functions. And in a companion paper (Boudreau and Knoblauch, 2013) we also study tradeoffs in utilitarian terms,[2] also both theoretically and with simulation experiments, though in that paper we focus on how the price of stability changes with ordinal categorizations of agents' preferences. Clearly, however, traditional utilitarianism is just one notion of welfare.

This brings us to our second issue. In order to present a formal, more general definition of the price of stability, we need a general method for assigning values to matchings. Many studies assign values to matchings, but this has always been done in an *ad hoc* manner. We 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.

As a brief and informal preview of our two key definitions, a social welfare function (SWF) assigns a non-negative real number to every ordered pair consisting of a matching and a preference profile. For a market of size n with preference profile P_n, the price of stability associated with an SWF f is the ratio of the maximum value of f over all matchings to the maximum value of f over all stable matchings. Then we will say there is a price tag attached to stability for a market of size n if that ratio is greater than one for some P_n. When defined as above, an SWF can 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.

In the next section, Section 2, we define SWFs over marriage matchings, provide several families of SWFs motivated by previous literature, and argue that SWFs in general are legitimate tools for the study of marriage matching. In Section 3 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, when markets are large, a randomly chosen preference profile will almost certainly show that the SWF in question comes with a price tag attached to stability. We also demonstrate that the price of stability for four of our examples is substantial. Section 4 concludes with a discussion of how simulation can provide information about the price of stability for many scenarios that are intractable via theory.

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.

2 In that paper (Boudreau and Knoblauch, 2013) we also consider another welfare measure, referred to as a Rawlsian measure, which is discussed below.

2. 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 n 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.[3] Man i's preferences are given by a one-to-one and onto ranking function $r_{m_i}: W \rightarrow \{1, 2, \ldots, n\}$ 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 preference 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 $\mu: M \rightarrow W$. Let $\mathcal{M}$ denote the set of all matchings. 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. But although a matching market's set of stable outcomes can be guaranteed non-empty, it is rarely single valued.

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 (Roth and Sotomayor, 1990, Theorem 3.8 attributed to Conway). 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. Let S denote the set of all stable matchings, and let μ_M and μ_W be the man-optimal and the woman-optimal matchings, 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 preference 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 is implicit in much of the existing work on matching.

Example 1.

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

Example 2.

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

3 As in the traditional version of the model (Gale and Shapley, 1962), we assume that all agents rank the option of remaining single last, so it is omitted for brevity.