

# **The High Cost of Stability in Two-Sided Matching: How Much Social Welfare Should be Sacrificed in the Pursuit of Stability?**

Robert L. Axtell  
George Mason University

Steven O. Kimbrough  
University of Pennsylvania

15 July 2008, WCSS

# Outline

1. Background on two-sided matching
2. Critique of received view
3. Agent model and results
4. Discussion

# 1. Background on two-sided matching

- No non-arbitrary definition of the class of problems. Basic idea: two sets of items,  $R$  and  $C$ . You want to match items from  $R$  with items from  $C$  and you have some measure of value or performance for the matching. How do you do it? Well, lots of special conditions and so forth. Instead of trying to say it all abstractly, we'll begin concretely with an example, then generalize it.
- That said, there are some established definitions, which we shall use.  
Remember: “the Marriage Problem” is really a huge abstraction of the marriage problem. We can study it without thinking that it somehow captures the essence of the underlying problem.

## The Marriage Problem (as traditionally described)

- There are  $m$  men and  $n$  women. Each man ranks each of the women with respect to his willingness or desire to marry her. Rank 1 is best, rank  $n$  ( $m$ ) is least preferred. The women similarly rank the men.
- How should we do the matching (of each of the  $m$  men with each of the  $n$  women)?
- We need a measure of performance (MOP), some way to assess how good a matching is. Then we can look for ways to achieve a matching that does well in terms of the measure of performance.

## Stability: One MOP (measure of performance)

- Suppose we have Bob, Carol, Ted and Alice. Bob and Carol are married. Ted and Alice are married. Bob prefers Alice as a wife to Carol. Alice prefers Bob as a husband to Ted. This situation is said to be unstable.
- Consider the preference orderings (rankings) shown in Figure 1.

	Carol	Alice
Bob	2, 1	1, 1
Ted	2, 2	1, 2

Figure 1: A sample array of 2-sided preference orderings

Marking marriages with  $\Rightarrow \dots \Leftarrow$

	Carol	Alice
Bob	$\Rightarrow 2, 1 \Leftarrow$	1, 1
Ted	2, 2	$\Rightarrow 1, 2 \Leftarrow$

Figure 2: Bob is married to Carol, Ted is married to Alice

Notice that under the present arrangements, Bob would prefer to be married to Alice and Alice would prefer being married to Bob. Instability!

NB: This is not a game in strategic form.

# Review of the Two-Sided Matching Problem and G-S [GS62]

In a two-sided matching problem, we are given two disjoint sets<sup>1</sup>  $R$  and  $C$  (think: Row and Column). Denote the elements of  $R$  as  $r_1, r_2, \dots, r_m$  and the elements of  $C$  as  $c_1, c_2, \dots, c_n$ .

**Definition 1. [Matched Pair]** *A matched pair is an ordered 2-tuple,  $\langle r, c \rangle$ , for which either*

- $r \in R$  and either  $(c \in C \text{ xor } r = c)$ , or
- $c \in C$  and either  $(r \in R \text{ xor } r = c)$ .

---

<sup>1</sup>See [GS62, page 12] for problems when the sets are not disjoint.

Under the intended interpretation if  $r \in R$  and  $c \in C$ , then  $\langle r, c \rangle$  is the match (e.g., marriage) of  $r$  and  $c$ . The definition, however, allows *self-matchings*, e.g.,  $\langle r, r \rangle$  and  $\langle c, c \rangle$ . A *restricted matched pair* is a matched pair formed under conditions that prohibit self matches. The interpretation of a self-match is that that particular self prefers not to be matched with any of the lower-valued possibilities. (E.g., would prefer to be unmarried.)

**Definition 2. [Match Array]** *A match array is an enumeration of all (legal) matched pairs.*

The following is a tabular representation in schematic form of a match array. Let  $|R| = m$  and  $|C| = n$ .

	$c_1$	$c_2$	$\dots$	$c_n$	$r_s$
$r_1$	$\langle r_1, c_1 \rangle$	$\langle r_1, c_2 \rangle$	$\dots$	$\langle r_1, c_n \rangle$	$\langle r_1, r_1 \rangle$
$r_2$	$\langle r_2, c_1 \rangle$	$\langle r_2, c_2 \rangle$	$\dots$	$\langle r_2, c_n \rangle$	$\langle r_2, r_2 \rangle$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$r_m$	$\langle r_m, c_1 \rangle$	$\langle r_m, c_2 \rangle$	$\dots$	$\langle r_m, c_n \rangle$	$\langle r_m, r_m \rangle$
$c_s$	$\langle c_1, c_1 \rangle$	$\langle c_2, c_2 \rangle$	$\dots$	$\langle c_n, c_n \rangle$	

Figure 3: Schematic representation in tabular format of a match array for  $R$  and  $C$

The row heading  $c_s$  indicates self matches for the corresponding

columns, and the column heading  $r_s$  indicates self matches for the corresponding rows. If matched pairs are unrestricted (i.e., self matching is permitted), the size of the match array is  $|R| \cdot |C| + |R| + |C|$ . Note that  $\langle c_s, r_s \rangle$  (the southeast corner of the table) is not permitted in the match array. If restricted matching is in force the size is  $|R| \cdot |C|$ . The *restricted match array* is the  $m \times n$  sub-table obtained from the match array by eliminating the column and row of self matches.

**Definition 3. [Match]** *A match (or matching)  $\mu$  is a set of matched pairs from a match array.*

Any individual member of  $R$  or  $C$  may have a *quota*, an upper bound on the number of matches in which it may participate. In marriage problems everyone characteristically has a quota of 1. In matching workers to firms, typically each worker has a quota of 1 and firms may have quotas greater than 1.

**Definition 4. [Feasible Match]** *A match  $\mu$  is a feasible match if no quota is violated.*

In terms of the match array, a row element is feasible if the number of matched pairs in the row is not greater than the row element's quota. Similarly a column element is feasible if the number of matched pairs in the column is less than or equal to its quota. A match is feasible if all of the row and column elements in its corresponding match array are feasible.

**Definition 5. [Simple Match]** *A simple match is a feasible match under the condition that each member of  $R$  or  $C$  has a quota of 1.*

A *valuation* for a row element  $r \in R$ ,  $V_r$ , is a preference ranking for  $r$  on  $C \cup r$ . Similarly for a valuation  $V_c$  of a column element  $c$ . We assume that ranking is strict (no ties) and stipulate that the most preferred object has a rank of 1 and the least preferred object the rank of  $|C|+1$  ( $|R|+1$ ). (In a restricted match, with no self-matching, the maximum rank values are  $|R|$  and  $|C|$ .) Denote the rank (preference index) of  $y$  for  $x$  by  $P_{xy}$  (mnemonic:  $P$  is *position* in the ranking, lower numbered being better). For each matched pair  $\langle x, y \rangle$  there is a corresponding *valuation pair*,  $\langle P_{xy}, P_{yx} \rangle$ . A *valuation array* (or “ranking matrix” [GS62, page 11]) is a match array with valuation pairs substituted for matched pairs.

	$c_1$	$c_2$	...	$c_n$	$r_s$
$r_1$	$\langle P_{r_1}c_1, P_{c_1}r_1 \rangle$	$\langle P_{r_1}c_2, P_{c_2}r_1 \rangle$	...	$\langle P_{r_1}c_n, P_{c_n}r_1 \rangle$	$\langle P_{r_1}r_1, P_{r_1}r_1 \rangle$
$r_2$	$\langle P_{r_2}c_1, P_{c_1}r_2 \rangle$	$\langle P_{r_2}c_2, P_{c_2}r_2 \rangle$	...	$\langle P_{r_2}c_n, P_{c_n}r_2 \rangle$	$\langle P_{r_2}r_2, P_{r_2}r_2 \rangle$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$r_m$	$\langle P_{r_m}c_1, P_{c_1}r_m \rangle$	$\langle P_{r_m}c_2, P_{c_2}r_m \rangle$	...	$\langle P_{r_m}c_n, P_{c_n}r_m \rangle$	$\langle P_{r_m}r_m, P_{r_m}r_m \rangle$
$c_s$	$\langle P_{c_1}c_1, P_{c_1}c_1 \rangle$	$\langle P_{c_2}c_2, P_{c_2}c_2 \rangle$	...	$\langle P_{c_n}c_n, P_{c_n}c_n \rangle$	

Table 1: Schematic representation in tabular format of a valuation array for  $R$  and  $C$

**Definition 6. [Acceptability]** *A match is **unacceptable** if it contains a matched pair for which  $P_{xy} > P_{xx}$ . A match is **acceptable** if it is not unacceptable.*

If  $x$  is matched to  $y$  in  $\mu$  (write:  $\mu(x) = y$ ) and  $x$  prefers to be self-matched to being matched to  $y$ , then  $\mu$  is unacceptable. Put otherwise, in an acceptable match, no one does worse than being self-matched.

**Definition 7. [Individually Rational Matching]** *A match  $\mu$  is individually rational if it is feasible and acceptable.*

Comment: this definition corresponds to and is equivalent to Definition 2.2 of Roth and Sotomayor [RS90, page 21].

We are now in position to be able to define the class of two-sided match problems, given  $R$  and  $C$  and the definitions above.

NB: Broader, alternative definitions are possible. This has thus a “technical” meaning.

**Definition 8. [Two-Sided Match Problem]** *A two-sided match problem is the problem of finding a high-quality, ideally optimal, individually rational matching between the two given sets  $R$  and  $C$ , and a given valuation array.*

Two important special cases of two-sided match problems are the marriage problem and the admissions problem.

**Definition 9. [Marriage Problem]** *A marriage problem is a two-sided match problem in which each member of  $R \cup C$  has a quota of 1.*

**Definition 10. [Admissions Problem]** *An admissions problem is a two-sided match problem in which at least one member of  $R$  ( $C$ , think: colleges) has a quota greater than 1 and all members of  $C$  ( $R$ , think: students) have a quota of 1.*

We will for the most part focus our attention on marriage problems.

Note that either problem may allow or disallow self-matches. By default we assume that self-matches are allowed and that in any matching if an agent is not matched to an agent of the opposite class (an  $r$  is matched to some member of  $C$ , and a  $c$  is matched to some member of  $R$ ), then the agent is matched to itself.

Obtaining any significant results for matching problems requires that “high-quality” in the definition of two-sided matching be specified. Of course, the notions of Pareto optimal matches (matches such that no other match exists that makes everyone better off) and Pareto dominant matches (matches such that in every other match everyone is strictly worse off) are available to us. The literature on two-sided matching has, however, focused on the concept of a stable match, which may be defined as follows.

**Definition 11. [Stable Match]** *A match  $\mu$  is **unstable** if it contains matched pairs  $\langle w, x \rangle, \langle y, z \rangle$  such that  $P_w x > P_w z$  and  $P_z y > P_z w$ . A match is **stable** if it is not unstable.*

Note: Valuations with lower rank values are preferred to valuations with higher rank values, so if  $P_w x > P_w z$ , then  $w$  prefers  $z$  over  $x$ . This definition corresponds to Gale and Shapley's definition [GS62, page 10] and to Definition 2.3 of Roth and Sotomayor [RS90, page 21].

Stability characterizes equilibrium in two-sided matching problems. Given a match there may well be many individuals with incentive to find different partners. Characterizing two-sided matching problems is the assumption that individuals cannot act alone. If  $r_i$  is unhappy with its partner in a match, we assume that  $r_i$  can do nothing unless there is a  $c_j$  that  $r_i$  prefers to its current partner, and  $c_j$  is less happy with its match than it would be paired with  $r_i$ . The marriage metaphor motivates the definition. If a man prefers the wife of his neighbor to his own wife, and

the neighbor's wife prefers the man to her own husband, we have an unstable situation. A matching is thus said to be stable if there is no pair of matched couples for which a pair of individuals has incentive to switch partners. Note that if a switch is made the jilted individuals may or may not benefit.

The literature on two-sided matching has largely been directed at finding stable matches. The seminal paper is [GS62], which presents an algorithms that are guaranteed to find stable matches for marriage problems and admissions problems. Originally referred to as the *deferred acceptance procedure*, the algorithms collectively has come to be called the Gale-Shapley (G-S) procedure. When necessary, we shall distinguish them by name:  $G-S_M$  for the marriage version (everyone has a quota of 1), and  $G-S_A$  for the admissions version. We will get to the G-S procedure shortly.

Gale and Shapley present two theorems in their paper.

**Theorem 1. [G-S Theorem 1 [GS62]]** *There always exists a stable set of marriages.*

That is, for the marriage problem, the  $G-S_M$  algorithm (deferred acceptance) will always find a stable match.

Assuming a definition of optimality—“A stable assignment is called *optimal* if every applicant [in an admissions problem; applicants have quotas of 1] is at least as well off under it as under any other stable assignment”—Theorem 2 asserts that  $G-S_A$  produces an optimal match.

**Theorem 2. [G-S Theorem 2 [GS62]]** *Every applicant is at least as well off under the assignment given by the deferred acceptance procedure as he would be under any other stable assignment.*

## Description of $G-S_M$

“To start, let each boy propose to his favorite girl. Each girl who receives more than one proposal rejects all but her favorite from among those who have proposed to her. However, she does not accept him yet, but keeps him on a string to allow for the possibility that someone better may come along later.

“We are now ready for the second stage. Those boys who were rejected now propose to their second choices. Each girl receiving proposals chooses her favorite from the group consisting of the new proposers and the boy on her string, if any. She rejects all the rest and again keeps the favorite in suspense.

“We proceed in the same manner. Those who are rejected at the

second stage propose to their next choices, and the girls again reject all but the best proposal they have had so far.

“Eventually (in fact, in at most  $n^2 - 2n + 2$  stages) every girl will have received a proposal, for as long as any girl has not been proposed to there will be rejections and new proposals, but since no boy can propose to the same girl more than once, every girl is sure to get a proposal in due time. As soon as the last girl gets her proposal the ‘courtship’ is declared over, and each girl is now required to accept the boy on her string.” [GS62, pages 12–3]

## 2. Critique of received view

- So what's not to like?
- There are anomalies or paradoxical results (here we proceed as philosophers)

## Anomaly #1: Sexual Bias (fairness)

G-S in terms of our match array representation, figure 3, is row-oriented. It matters whether men are assigned to the set  $R$  or whether women are. To illustrate, on page 11 they present their *Example 1*, in the form of a valuation array for a  $3 \times 3$  marriage problem.

*Example 1.* The following is the “ranking matrix” [valuation array] of three men,  $\alpha$ ,  $\beta$ , and  $\gamma$ , and three women,  $A$ ,  $B$ , and  $C$ .

	$A$	$B$	$C$
$\alpha$	1,3	2,2	3,1
$\beta$	3,1	1,3	2,2
$\gamma$	2,2	3,1	1,3

The first number of each pair in the matrix gives the ranking of women by the men, the second number is the ranking of the men by the women. Thus,  $\alpha$  ranks  $A$  first,  $B$  second,  $C$  third, while  $A$  ranks  $\beta$  first,  $\gamma$  second, and  $\alpha$  third, etc.

There are six possible sets of marriages; of these, three are stable. One of these is realized by giving each man his first choice, thus  $\alpha$  marries  $A$ ,  $\beta$  marries  $B$ , and  $\gamma$  marries  $C$ . Note that although each woman gets her last choice, the arrangement is nevertheless stable. Alternatively one may let the women have their first choices and marry  $\alpha$  to  $C$ ,  $\beta$  to  $A$ , and  $\gamma$  to  $B$ . The third stable arrangement is to give everyone his or her second choice and have  $\alpha$  marry  $B$ ,  $\beta$  marry  $C$ , and  $\gamma$  marry  $A$ . The reader will easily verify that all other arrangements are unstable.

Transparently, in this example, for whichever stable match of the three

stable matches is found by the G-S it will be the case that there is another stable match in which someone does worse than in the match found by G-S. If G-S finds the first match, then all the men do worse in matches 2 and 3. If G-S finds the second match, then all the men fare better in the first and third matches. If G-S finds the third match, then all the men do better in the first match. Optimality is defined by G-S as Pareto dominance. There is here no Pareto dominant solution among the stable matches.

If the men are assigned to  $R$  in this example, then G-S finds the match in which each man gets his first choice. Contrariwise, if the women get the the  $R$  rôle, the G-S finds the match in which each woman gets here first choice. G-S cannot find the stable match in which everyone gets his or her second choice. This, as shown in [GS62], is not a peculiarity of the example. Under the  $G-S_M$  procedure, whichever group gets assigned to be  $R$  obtains a matching that is optimal for it in

the following sense:

**Definition 12. [From [GS62]]** *A stable assignment is called **optimal** if every applicant is at least as well off under it as under any other stable assignment.*

See also Theorem 2.12 of [RS90, page 32]. And in general (assuming throughout strict preferences), the side assigned  $R$  gets its optimal stable matching, while the side assigned  $C$  gets its pessimal stable matching; each member of  $C$  gets its least preferred achievable partner ([RS90, page 33], [Knu76]).

## Anomaly #2: G-S Theorem 2 Revisited (fairness)

The sexual bias results for  $G-S_M$  raise the question of social welfare. Under G-S one class is assigned to be  $R$  and in consequence gets results that are, in the sense defined, optimal for it. The other class gets its worst deal. Although there may exist compromise matches, they are not available via G-S, which is also silent on which class to favor.

Theorem 2 of [GS62], the proof of optimality for the  $G-S_A$  algorithm (deferred acceptance) for the admissions problem, might seem to avoid the social welfare question. It is important to understand that it does not. The algorithm simply *defines* the applicants to be the favored class, at the expense of the universities. To illustrate, consider the following problem. There are two schools, Hard University and Knocks College, and each has a quota of 2 applicants. There are four applicants, Bob,

Ted, Carol, and Alice. The valuation array (G-S's ranking matrix) is as in figure 4.

	Hard	Knocks
Bob	2,1	1,4
Ted	1,2	2,3
Carol	2,3	1,2
Alice	1,4	2,1

Figure 4: Valuation array for an admissions problem

As always we have two cases for the G-S algorithm, depending on which side is assigned  $R$  (and hence does the proposing) and which  $C$ . Here is the G-S description of deferred acceptance for the admissions problem when the students propose (are Row players) [GS62, page 13]:

First, all students apply to the college of their first choice. A college

with a quota of  $q$  then places on its waiting list the  $q$  applicants who rank highest, or all applicants if there are fewer than  $q$ , and rejects the rest. Rejected applicants then apply to their second choice and again each college selects the top  $q$  from among the new applicants and those on its waiting list, puts these on its new waiting list, and rejects the rest. The procedure terminates when every applicant is either on a waiting list or has been rejected by every college to which he is willing and permitted to apply. At this point each college admits everyone on its waiting list and the stable assignment has been achieved.

Case 1: Students propose. In the example to hand, in the first round Bob and Carol apply to Knocks, and Ted and Alice apply to Hard. At this point the procedure terminates. Figure 5 shows the pairings; it is obvious the match is stable.

	Hard	Knocks
Bob		1,4
Ted	1,2	
Carol		1,2
Alice	1,4	

Figure 5: Stable match with students proposing

What happens in case 2, with the colleges proposing? Gale and Shapley [GS62] simply do not propose an algorithm for admissions with the colleges proposing. G-S for admissions problems stipulates that students, not colleges, propose. Can there be stable matches other than the one found by G-S? Figure 6 is one for the example to hand.

	Hard	Knocks
Bob	2,1	
Ted	1,2	
Carol		1,2
Alice		2,1

Figure 6: Second stable matching for the example

Is the matching of figure 6 obtainable by a close analog of  $G-S_A$ , but with colleges proposing (assigned to  $R$ )? Consider the following procedure.

**Procedure 1. [Colleges Proposing Admissions]** *All colleges propose to their highest-ranked students. A student with more than one proposal rejects all but its highest-ranked proposal. Any college with a presently unrejected proposal adds the student to its waiting list, decrementing*

*its quota. Every college with a positive quota remaining proposes to its next most preferred student. Every student with more than one proposal rejects all but its highest-ranked proposal. Every college now updates its quota to reflect the number of its accepted proposals. The procedure terminates when every college has proposed to each of its ranked students, or before if stopping will not change the outcome.*

We can now describe case 2, using the Colleges Proposing Admissions procedure.

Case 2: Schools propose. Hard proposes to Bob and Knocks proposes to Alice. There are no conflicts so round 2 ensues. Hard proposes to Ted and Knocks proposes to Carol. Again, there are no conflicts. The procedure terminates because both schools have proposed to all of their ranked students. The resulting match is that of figure 6.

As you may easily verify, the Colleges Proposing Admissions is directly analogous to the  $G-S_A$  procedure, and generally shares its properties, with the exception that the colleges get their optimal match.

## Anomaly #3: Social Welfare and Unstable Matches

**Definition 13. [Summation Social Welfare Function]** *In a two-sided matching problem, with sides  $X$  and  $Y$ , the summation social welfare of a matching,  $\mu$ , is the sum total of rank points realized by each side in  $\mu$ .*

**Definition 14. [SumProduct Social Welfare Function]** *In a two-sided matching problem, with sides  $X$  and  $Y$ , the sumproduct social welfare of a matching,  $\mu$ , is the sum total of the products of rank points of the pairs that are matched in  $\mu$ .*

**Proposition 1. [Existence of Socially Superior Unstable Matches]** *In the two-sided matching problem with  $|X| > 3$  and  $|Y| > 3$ , there exist individual preference orderings such that there are stable matches that are socially inferior to unstable matches, under the summation and sumproduct social welfare functions.*

# Summing Up

There are other anomalies/problems as well. Key for us is to note that there are (at least) three criteria by which to evaluate a matching:

1. Stability (G-S et al., [RS90])
2. Social welfare (NB linear programming formulation)
3. Fairness (e.g., [KK06], [FNK06])

Stable matchings under G-S do not maximize social welfare and do not lead to fair outcomes.

### 3. Agent model and results

The challenge:

How might we find matchings that either (a) are stable and are improved wrt social welfare and/or fairness, or (b) are not stable, but not unacceptably so, and are improved wrt social welfare and/or fairness?

How, in particular, might we build agent-based models to investigate the problems of two-sided matching?

NB: Real-world complications often vex the deferred-acceptance approach.

Read the paper, but briefly. . .

## Agent Model: Distributed Matching

- Two types of agents.
- Each individual of each type randomly (uniformly) ranks each individual of the other type.
- Each step a small number of agents, randomly chosen, become active and make proposals of engagement that are accepted or not.
- Steps accumulate into periods, when the number of agents activated equals the number of agents.
- With some probability, an engaged pair get married.

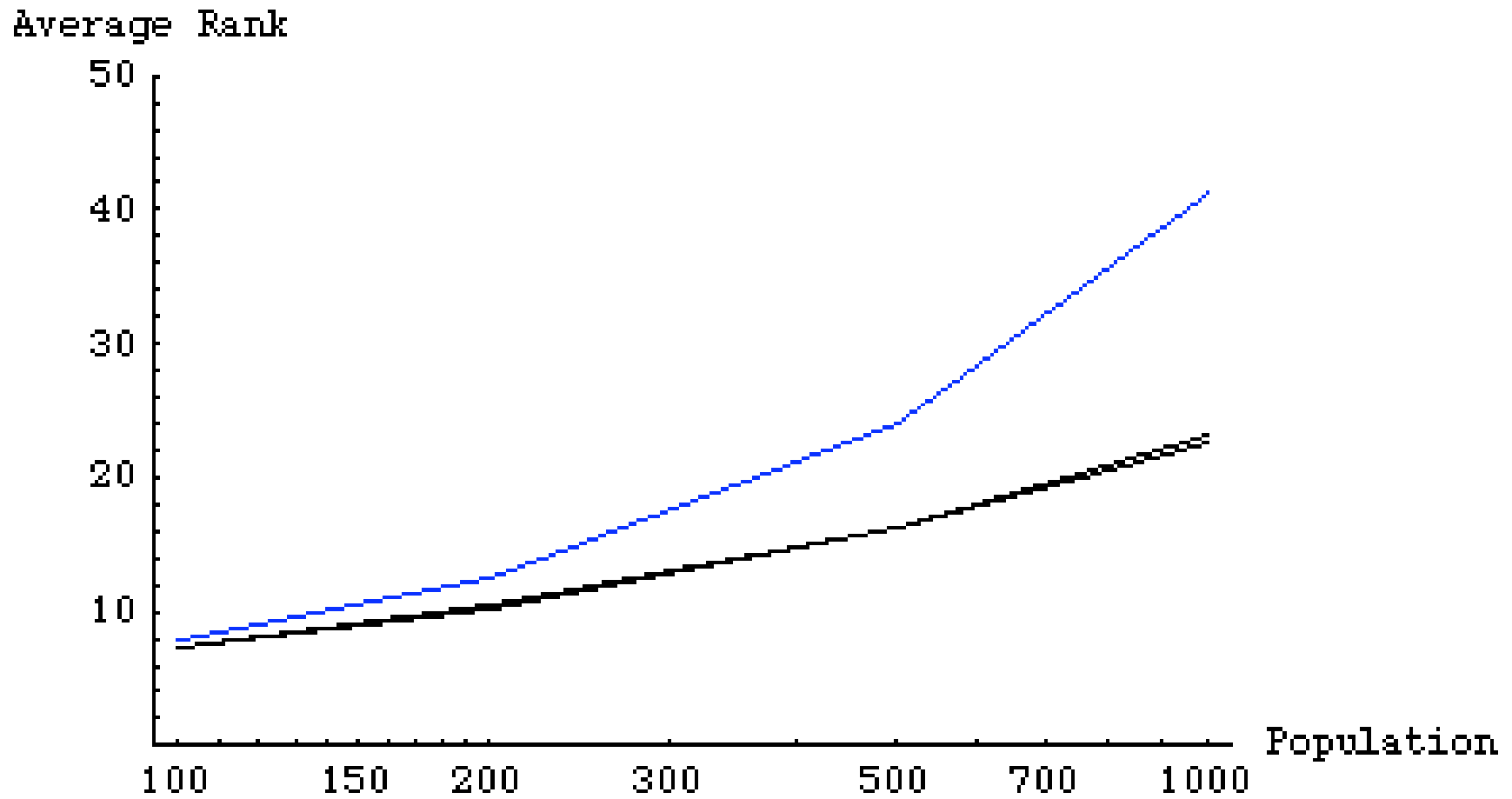


Figure 7: Figure 4 from paper. Compared to G-S (blue), out agents are better on SW and on fairness.

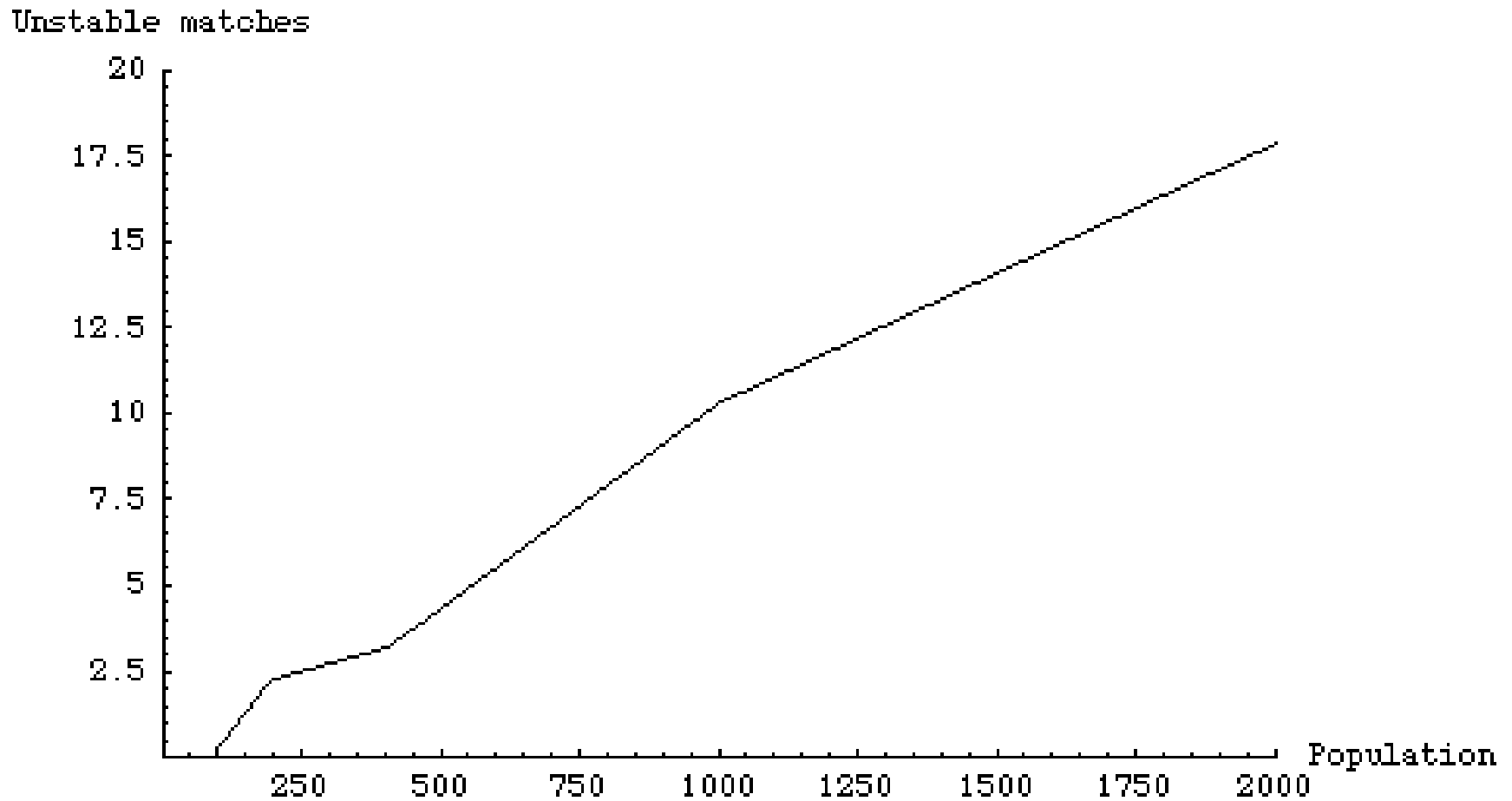


Figure 8: Figure 8 from paper. Slow increase in the number of unstable matches. Uniform random preferences, 100 periods, 1% acceptance.

## 4. Discussion

1. See the paper for more results.
2. In a nutshell: Distributed Matching is more realistic, fairer and has superior SW.
3. And can be adapted to more general settings.
4. We marvel at the strange attractiveness of equilibrium concepts.

## References

- [FNK06] Tomoko Fuku, Akira Namatame, and Taisei Kaizouji, *Collective efficiency in two-sided matching*, Artificial Economics: Agent-Based Methods in Finance, Game Theory and Their Applications (Philippe Mathieu, Bruno Beaufils, and Olivier Brandouy, eds.), Springer-Verlag, Berlin, Germany, 2006, pp. 115–126.
- [GS62] D. Gale and L. S. Shapley, *College admissions and the stability of marriage*, The American Mathematical Monthly **69** (1962), no. 1, 9–15.
- [KK06] Bettina Klaus and Flip Klijn, *Procedurally fair and stable matching*, Economic Theory **27** (2006), 431–447.

- [Knu76] Donald E. Knuth, *Marriages stables*, Les Presses de l'Université de Montreal, Montreal, Canada, 1976.
- [RS90] Alvin E. Roth and Marilda A. Oliveira Sotomayor, *Two-sided matching: A study in game-theoretic modeling and analysis*, Cambridge University Press, Cambridge, United Kingdom, 1990.

# End Note

\$Id: twosidedmatching-foils.tex 381 2008-07-15 17:17:49Z sok \$