Games and Economic Behavior 123 (2020) 68–80

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Evolution and Rawlsian social choice in matching ☆

Ennio Bilancini [a], Leonardo Boncinelli [b,*], Jonathan Newton [c]

[a] IMT School of Advanced Studies Lucca, Piazza S. Francesco 19, 55100 Lucca, Italy
[b] Dipartimento di Scienze per l'Economia e l'Impresa, University of Florence, Via delle Pandette 9, 50127 Firenze, Italy
[c] Institute of Economic Research, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

ARTICLE INFO

Article history:
Received 9 January 2020
Available online 1 July 2020

JEL classification:
C71
C72
C73
C78
D71

Keywords:
Evolution
Stochastic stability
Matching
Rawlsian

ABSTRACT

This paper considers the marriage problem under dynamic rematching. It is shown that if players who obtain higher payoffs are less likely to experiment with non-best response behavior, then matchings selected in the long run will belong to the set of Rawlsian stable matchings – the set of stable matchings which maximize the payoff of the worst off player. Conversely, alternative behavioral rules will fail to select Rawlsian stable matchings in some environments. This constitutes an evolutionary axiomatization of Rawlsian stable matchings in terms of the behavioral rules that give rise to them.

1. Introduction

It is common in evolutionary economics to assume that agents follow explicit behavioral rules. From this, implications for society are derived. For example, in the classic marriage model of two sided matching (Gale and Shapley, 1962), play under pairwise best response dynamics converges to the core (Roth and Vande Vate, 1990). In two strategy coordination games, play under certain perturbed best response dynamics (Young, 1993; Kandori et al., 1993) converges to the risk dominant Nash equilibrium. Such models can be interpreted as giving sufficient conditions on underlying behavior for the emergence of a given social norm.

Necessary conditions on behavior for the emergence of a social norm are not usually given. Striking exceptions to this are Blume (2003) and Maruta (2002), who give necessary and sufficient conditions on perturbed individualistic best response dynamics for convergence to the risk dominant Nash equilibrium of a two strategy coordination game played in a population. These results can be interpreted as behavioral axiomatizations of the form: "Solution X will emerge from a dynamic process of strategy updating if and only if conditions Y on behavior are satisfied." Such an axiomatic approach provides a rigorous and comprehensive answer to how robust any given social norm is to misspecification of the behavioral rule, thus providing a context-specific investigation of the Bergin and Lipman (1996) critique that the identity of emergent norms may be non-robust to variations in the behavioral rule.

☆ We thank Bettina Klaus for advice on related matching literature and Yuval Heller for a useful discussion.
* Corresponding author.
E-mail addresses: ennio.bilancini@imtlucca.it (E. Bilancini), leonardo.boncinelli@unifi.it (L. Boncinelli), newton@kier.kyoto-u.ac.jp (J. Newton).

https://doi.org/10.1016/j.geb.2020.06.004
0899-8256/

In the current paper, we follow this approach and apply it to the marriage model of two sided matching. Specifically, we characterize *Rawlsian stable matchings* – stable matchings which maximize the payoff of the worst off player (see Rawls, 1971; Masarani and Gokturk, 1989), in terms of the behavior that leads to the emergence of such matchings under an evolutionary process.

Consider pairwise dynamics that differ from Roth and Vande Vate (1990) in that there is some small probability of players choosing non-best responses. When the propensity of a player to play a non-best response decreases in his current payoff, almost to the exclusion of other considerations, we say that he follows a *condition dependent* behavioral rule (Bilancini and Boncinelli, 2020).[1] An interpretation is that players who earn high payoffs are less inclined to experimentally change their behavior in search of higher payoffs.

Our 'sufficiency' theorem (Theorem 1) shows that Rawlsian stable matchings emerge when players follow condition dependent behavioral rules. Specifically, Rawlsian stable matchings are *stochastically stable* (Foster and Young, 1990). This result, while interesting, is a mostly straightforward application of the one-shot deviation principle for stability in matching problems (Newton and Sawa, 2015).

Our 'necessity' theorem (Theorem 2) shows that condition dependence is necessary to guarantee Rawlsian stability. That is, every rule that is not condition dependent will fail to select Rawlsian stable matchings in some environment. This is the more novel aspect of the paper, and is proved constructively. Given a behavioral rule that is not condition dependent, and given sets of men and women on either side of the matching market, we find payoffs under which there exists a stochastically stable matching which is not Rawlsian.

A germane comparison is to traditional axiomatic approaches. These express solution concepts in terms of the satisfaction of some desirable property (e.g. Pareto efficiency, monotonicity) of the solution itself. That is, the axioms are restrictions placed directly on the solution. The approach of the current paper is to express a solution concept (Rawlsian stable matching) in terms of the behavioral rules that give rise to it (condition dependent rules). That is, the axioms are restrictions placed on behavior.

It is worth noting that an appealing behavioral rule will not necessarily lead to a solution concept with intuitively appealing properties. For example, Hwang et al. (2018a) show that the logit choice rule gives rise to a bargaining solution that fails to satisfy many of the appealing properties satisfied by standard bargaining solutions.[2] In the current paper both the behavioral rule and the emergent social norm have independent appeal. Condition dependent rules can be understood as abiding by a maxim to "focus on what you have, not on what you do not have", or in the words of the Decalogue, "thou shalt not covet". Regarding the social norm, Rawls (1971) gives a philosophical argument in favor of structuring institutions to maximize the wellbeing of the worst off in society, subject to the restraints of maintaining a liberal society and due consideration for future generations.

The paper is organized as follows. Section 2 discusses some related literature. Section 3 gives the model. Section 4 gives the main results. Section 6 concludes. Ancillary results and proofs are relegated to the appendix.

2. Related literature

There is a significant literature that considers the emergence of norms in matching problems. For matching problems with non-transferable utility, pairwise best response (Roth and Vande Vate, 1990) and pairwise best response with uniform deviations (Jackson and Watts, 2002) select the set of stable matchings. Newton and Sawa (2015) give general methodology for such models and show that coalitional logit choice (Sawa, 2014) selects a non-transferable utility version of the least core (Maschler et al., 1979).[3]

Similarly, for matching with transferable utility, it has been shown that pairwise best response (Chen et al., 2016; Biró et al., 2013; Klaus and Payot, 2015; Nax et al., 2013) and pairwise best response with uniform deviations (Klaus and Newton, 2016) select the core (Gillies, 1959); coalitional logit selects the least core (Nax and Pradelski, 2015); and a logit-like class of weakly payoff dependent processes will select the interior core if there is a unique optimal matching, but may further restrict selection when there are multiple optimal matchings (Klaus and Newton, 2016).

The above papers are all part of the Evolutionary Nash Program that considers links between evolutionary game theory and cooperative game theory. Furthermore, in considering pairwise behavioral rules, they are part of the literature on evolution and collective agency (see Newton, 2018, for a survey covering both of these literatures).

The most well known axiomatizations of solution concepts are probably those of the Nash (1950), Kalai and Smorodinsky (1975) and Egalitarian (Kalai, 1977) bargaining solutions. However, there has been work on the axiomatization of solutions to problems of matching and object allocation. For example, axiomatizations have been given for allocation rules such as

[1] This terminology follows from the evolutionary biology literature suggesting that organisms in a worse 'condition' have a higher mutation rate (Agrawal, 2002).

[2] For an overview of the logit choice rule, see Alós-Ferrer and Netzer (2010). More generally, logit choice is a member of the classes of *skew-symmetric* rules (Blume, 2003), *payoff-based* rules (Peski, 2010) and *payoff-difference based* rules (Newton, 2019). These classes share the common feature that the probability of a deviation decreases in the payoff loss to oneself, and in the latter case possibly also to others, from the deviation in question.

[3] Pairwise best response (Roth and Vande Vate, 1990) is a natural extension of best response (Cournot, 1838). Pairwise best response with uniform deviations (Jackson and Watts, 2002) is a natural extension of best response with uniform perturbations (Young, 1993; Kandori et al., 1993). Coalitional logit choice (Sawa, 2014) is a natural extension of logit choice (e.g. Blume, 1993).

deferred acceptance (Kojima and Manea, 2010; Ehlers and Klaus, 2016), immediate acceptance (Doğan and Klaus, 2018; Kojima and Ünver, 2014), top trading cycles (Ma, 1994) and serial dictatorship (Svensson, 1999).

Finally, we note that recent experimental evidence shows that behavior in the lab is highly consistent with the myopic best response rule in simple coordination games, and that deviations from such rules vary with patterns of realized and expected payoffs (Mäs and Nax, 2016; Lim and Neary, 2016; Hwang et al., 2018b). In at least one case (Mäs and Nax, 2016) deviations are best explained by the condition dependent model (Bilancini et al., 2020).

3. Model

3.1. The marriage problem

We follow the description of the marriage problem in Jackson and Watts (2002). There is a set of players, N, which is divided into a set of men, $M = \{m_1, \ldots, m_k\}$, with $k \geq 2$, and a set of women, $W = \{w_1, \ldots, w_l\}$, with $l \geq 2$. An undirected network g is a set of edges, each edge comprising a pair of players $\{i, j\} \subset N$, $i \neq j$. We denote $\{i, j\}$ by ij so that $ij \in g$ indicates that there is an edge between players i and j in network g. Let $\mathcal{G}$ denote the set of all undirected networks on N. Let $g(i) = \{j : ij \in g\}$ denote the set of players linked to player i in network g. $g(i) = \emptyset$ means that i is single in g. The set of matchings in the marriage problem, G, is the set of undirected networks in which each woman is linked to at most one man, and each man is linked to at most one woman:

$$G = \{g \in \mathcal{G} : \ (\forall ij \in g, \ i \in M \Leftrightarrow j \in W), \ (\forall i \in N, \ |g(i)| \leq 1)\}.$$

In a slight abuse of notation, we sometimes write $g(i) = j$ for $g(i) = \{j\}$.

The vector of utilities obtained from network g by the players is given by $u : G \to \mathbb{R}^{|N|}$. Player i obtains payoff $u_i(g)$ from network g, and this payoff depends only on the match of i. That is, for each i, $u_i(g) = u_i(g')$ if $g(i) = g'(i)$. We assume that players are never indifferent between two potential matches: $g(i) \neq g'(i)$ implies that $u_i(g) \neq u_i(g')$. Therefore, if $g(i) \neq \emptyset$, then $u_i(g) = u_i(\{ig(i)\})$, and if $g(i) = \emptyset$, then $u_i(g) = u_i(\emptyset)$. We denote by $\mathcal{U}$ the set of all possible vectors of utilities u. Define

$$g - ij := g \setminus \{ij\}$$

as the network g with the edge ij removed if it exists in g. Similarly, define

$$g + ij := g \setminus \left\{ kl : \left(k = i \text{ and } l \in g(i) \right) \text{ or } \left(k = j \text{ and } l \in g(j) \right) \right\} \cup \{ij\}$$

as the network g with the edge ij added and any existing edges exiting i and j removed.

Definition 1. A matching $g \in G$ is (pairwise) stable if:

(i) $\forall ij \in g, \quad u_i(g) > u_i(g - ij)$.
(ii) $\nexists i \in M, \ j \in W : \quad u_i(g + ij) > u_i(g)$ and $u_j(g + ij) > u_j(g)$.

We denote the set of stable matchings by S. The set of stable matchings corresponds to the core of the problem: the set of matchings from which no subset of players can improve their payoffs by removing and adding edges in a coordinated manner.

A stable matching is *Rawlsian* if it is amongst the stable matchings that maximize the lowest payoff amongst all players. We denote the set of Rawlsian stable matchings by Ra.[4]

Definition 2 *(Rawlsian stable matchings).*

$$Ra = \arg \max_{g \in S} \min_{i \in N} u_i(g).$$

Example 1 *(Stable matchings and Rawlsian stable matchings).* Let $M = \{m_1, m_2\}$, $W = \{w_1, w_2\}$, and players' payoffs from a given match be given by the matrix in Fig. 1. Payoffs from remaining single are assumed to be zero. The stable matchings are $g_M = \{m_1 w_1, m_2 w_2\}$ and $g_W = \{m_1 w_2, m_2 w_1\}$. g_M is the man-optimal matching and g_W the woman-optimal matching. Note that $\min_{i \in N} u_i(g_M) = u_{w_1}(g_M) = 2$ and $\min_{i \in N} u_i(g_W) = u_{m_2}(g_W) = 5$, therefore $Ra = \{g_W\}$.

[4] Masarani and Gokturk (1989) defines a version of such matchings, which they call *stably fair*, for an environment with ordinal preference rankings. This corresponds to our definition if we let the payoff from being matched to one's kth favorite partner be equal to $1/k$.

	w_1	w_2
m_1	8, 2	7, 9
m_2	5, 6	7, 5

Fig. 1. Payoffs for matchings in Example 1. Entries give payoffs for row and column players respectively from being matched to one another. For example, m_1 obtains a payoff of 7 when matched to w_2.

3.2. Dynamic rematching

We follow the dynamic process of rematching in Newton and Sawa (2015). Pairs of players meet and can decide whether to dissolve an existing partnership or create a new partnership. A player will usually only agree to a change of partner when this leads to a higher payoff for him or herself. However, from time to time, players make mistakes and take actions which reduce their payoffs, whether it be leaving or creating a partnership.

We consider a family of dynamic processes $\{P_\eta\}_\eta$ parameterized by $\eta \in [0, \bar{\eta})$. Let g^t be the network in period t. At the beginning of period $t+1$, a pair of players (i, j) is selected at random according to a distribution $F_{g^t}(.)$ with full support on $\{(i, j) : i \in M, j \in W\}$. Let g^{t+1} be determined as follows:

If i and j are currently matched together, they consider separating, and if they are not currently matched together, they consider forming a partnership. That is, they consider a move to the matching g' given by

$$g' = \begin{cases} g^t - ij & \text{if } g^t(i) = j, \\ g^t + ij & \text{if } g^t(i) \neq j. \end{cases} \tag{3.1}$$

Each of i and j independently accept the proposed change with probabilities $A^i_\eta(g^t, g')$ and $A^j_\eta(g^t, g')$ respectively.[5]

For $\eta = 0$, we refer to the process as the *unperturbed dynamic*. We assume that, for all $k \in N$, $A^k_0(g^t, g') > 0$ if and only if $u_k(g') \geq u_k(g)$. That is, under the unperturbed dynamic, a player accepts a change if and only if he or she would not lose payoff from the change in question. For $\eta > 0$, we refer to the process as the *perturbed dynamic*. Under these dynamics, a player will, with positive probability, make a *deviation* and agree to a change that would reduce his or her payoff. We assume that $A^k_\eta(g^t, g')$ is continuous in η and strictly positive for all $\eta > 0$.

If i and j are currently together, it only takes one of them to force a separation. That is, if $g^t(i) = j$, then

$$g^{t+1} = \begin{cases} g^t - ij & \text{with prob. } 1 - \left(1 - A^i_\eta(g^t, g')\right)\left(1 - A^j_\eta(g^t, g')\right), \\ g^t & \text{with prob. } \left(1 - A^i_\eta(g^t, g')\right)\left(1 - A^j_\eta(g^t, g')\right). \end{cases} \tag{3.2}$$

If i and j are not currently together, they must both agree in order to form a new partnership. That is, if $g^t(i) \neq j$, then

$$g^{t+1} = \begin{cases} g^t + ij & \text{with prob. } A^i_\eta(g^t, g')\, A^j_\eta(g^t, g'), \\ g^t & \text{with prob. } 1 - A^i_\eta(g^t, g')\, A^j_\eta(g^t, g'). \end{cases} \tag{3.3}$$

The unperturbed dynamic is essentially identical to the dynamic of Roth and Vande Vate (1990). Our results concern perturbations of such a dynamic. In particular, we study stationary distributions of such perturbed processes for small values of η. Therefore, we are interested in the exponential decay rates of acceptance probabilities as $\eta \to 0$. We let these decay rates be given by a function $\varphi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ of the payoffs of the current matching and the proposed matching,

$$\lim_{\eta \to 0} -\eta \log A^k_\eta\left(g^t, g'\right) = \varphi\left(u_k(g^t), u_k(g')\right) \quad \text{for all} \quad k \in N. \tag{3.4}$$

Note that $u_k(g^t) \leq u_k(g')$ implies that $A^k_\eta(g^t, g')$ converges to $A^k_0(g^t, g') > 0$ as $\eta \to 0$, therefore $\varphi(u_k(g^t), u_k(g')) = 0$. If $u_k(g^t) > u_k(g')$, then $A^k_\eta(g^t, g')$ converges to $A^k_0(g^t, g') = 0$ and we assume that this decay is at an exponential rate. That is, $\varphi(u_k(g^t), u_k(g')) > 0$.

Examples of strategy revision rules that satisfy our assumptions include best response with uniform deviations, applied to the marriage problem by Jackson and Watts (2002), which gives

$$\varphi\left(u_k(g), u_k(g')\right) = \begin{cases} 0 & \text{if } u_k(g) \leq u_k(g'), \\ 1 & \text{if } u_k(g) > u_k(g'), \end{cases} \tag{3.5}$$

and the logit choice rule, applied to the marriage problem by Newton and Sawa (2015), which gives

[5] This assumption of independent acceptance of a proposed change is not present in Newton and Sawa (2015). The model can be considered without this assumption, but the current presentation is more in keeping with typical practice in the field. Notably, it combines the class of rules considered by Maruta (2002) with the accept-reject form of coalitional decision making used by Sawa (2014) in defining coalitional logit choice.

$$\varphi\left(u_k(g), u_k(g')\right) = \max\left\{0, u_k(g) - u_k(g')\right\}. \tag{3.6}$$

If, conditional on $u > u'$, $\varphi(u, u')$ only depends on its first argument, then Maruta (2002) refers to the rule as *purely aspirational*. Bilancini and Boncinelli (2020) refer to such rules as *condition dependent*.[6] The definition we use here allows φ to depend on its second argument, but only when the first argument is constant.

Definition 3. Behavior is *condition dependent* if φ is such that, for all $u, u', v, v' \in \mathbb{R}$, $u > u'$, $v > v'$, $u > v$, we have that $\varphi(u, u') > \varphi(v, v')$.

That is, under condition dependence, a strictly higher current payoff leads to a strictly higher φ (when φ is positive), implying that acceptance of a detrimental change is less likely when current payoffs are higher.

3.3. Stationary distribution and stochastic stability

As a chain with $\eta > 0$ is irreducible, there exists a unique stationary distribution π_η. For convenience, we assume the following.[7]

Assumption 1 *(existence of limit).*

$$\pi_0 := \lim_{\eta \to 0} \pi_\eta \quad \text{exists.}$$

A matching g is *stochastically stable* if $\pi_0(g) > 0$. This implies that there must exist at least one such matching. We denote the set of stochastically stable matchings by SS.

Definition 4 *(stochastically stable matchings).*

$$SS := \{g \in G : \pi_0(g) > 0\}.$$

All stochastically stable matchings belong to recurrent classes of the unperturbed process (Young, 1993) and from any matching there exists a finite sequence of transitions under the unperturbed process that culminates in a stable matching being reached (Roth and Vande Vate, 1990; Jackson and Watts, 2002). Therefore, a set of states is a recurrent class of the unperturbed process if and only if it contains a single stable matching and no other states. Consequently, $SS \subseteq S$. The identity of the stochastically stable matchings is important, as for small error probabilities the process will spend almost all of the time at these matchings.

4. Results

Our first result is that condition dependence leads to Rawlsian stable matchings.

Theorem 1. *If behavior is condition dependent, then* $\forall u \in \mathcal{U}$*, we have* $SS \subseteq Ra$.

The proofs of our theorems require additional notation, concepts and results given in Appendix A. Here, we explain the logic behind Theorem 1 using Example 1. Recall that lower values of φ correspond to more probable transitions.

Step 1: The stable matchings from which transitions to another matching are least probable are a subset of the Rawlsian stable matchings.

By definition of a stable matching, any transition from a stable matching to another matching will involve at least one player losing payoff. Under condition dependence, the most probable such transition involves a player whose current payoff is lowest. At g_M, this player is w_1 and the transition in which w_1 leaves m_1 to become single has decay rate

$$\varphi\left(u_{w_1}(g_M), u_{w_1}(\emptyset)\right) = \varphi(2, 0).$$

At g_W, this player is m_2 and the transition in which m_2 leaves w_1 to become single has decay rate

$$\varphi\left(u_{m_2}(g_W), u_{m_2}(\emptyset)\right) = \varphi(5, 0).$$

[6] Maruta (2002) and Bilancini and Boncinelli (2020) consider coordination games played by members of a population against all other members. In Maruta (2002) deviation probabilities depend on the current average payoff from playing against all other players. In Bilancini and Boncinelli (2020) deviation probabilities depend on the realized payoff after being randomly matched with one other player. In the current paper, each player has a maximum of a single partner, so these approaches are identical.

[7] This condition could be avoided if stochastically stable states were defined as states for which $\pi_\eta(.) \nrightarrow 0$ as $\eta \to 0$.

	w_1	w_2
m_1	v, z	z, u
m_2	z, u	u, z

Fig. 2. Payoffs for matchings to illustrate the logic of the proof of Theorem 2. Entries give payoffs for row and column players respectively from being matched to one another. Assume that $z > u > v$.

Definition 3 implies that $\varphi(2,0) < \varphi(5,0)$, so it is easier to leave g_M than it is to leave g_W. This is a consequence of the lowest payoff of any player at g_M (i.e., 2) being lower than the lowest payoff of any player at g_W (i.e., 5). That is, $g_M \notin Ra$ and $g_W \in Ra$.

Step 2: Stochastically stable matchings are a subset of the stable matchings from which transitions to another matching are least probable.

This is the one-shot deviation principle of Newton and Sawa (2015), a formal statement of which can be found in Appendix A.

Combining Steps 1 and 2, it must be that $SS = \{g_W\} = Ra$ in accordance with the statement of Theorem 1. Note that the argument given above is simplified as there are no single players at the stable matchings in Example 1. When there are single players at stable matchings, the argument is more involved and requires a series of lemmas given in Appendix B.

Our second result is that if behavior is not condition dependent, then there exist situations in which there are stochastically stable matchings which are not Rawlsian stable matchings.

Theorem 2. *If behavior is not condition dependent, then* $\exists u \in \mathcal{U}$ *such that* $SS \nsubseteq Ra$.

We again describe the logic behind the theorem.

Step 1: Construct an example with two men and two women for which $SS \nsubseteq Ra$**.**

If behavior is not condition dependent, then Definition 3 implies that there exist $u, u', v, v' \in \mathbb{R}$, $u > u'$, $v > v'$, $u > v$, such that $\varphi(u,u') \leq \varphi(v,v')$.

Let $M = \{m_1, m_2\}$, $W = \{w_1, w_2\}$, and players' payoffs from a given match be given by the matrix in Fig. 2. Payoffs from remaining single are v' for m_1 and u' for m_2, w_1, w_2. The stable matchings are $g_W = \{m_1w_1, m_2w_2\}$ and $g_M = \{m_1w_2, m_2w_1\}$. g_W is the woman-optimal matching and g_M the man-optimal matching. Note that $\min_{i\in N} u_i(g_W) = v$ and $\min_{i\in N} u_i(g_M) = u$, therefore $Ra = \{g_M\}$.

From g_W, there are the following types of transition to another matching. m_1 could accept a proposed separation from w_1. This would change his payoff from v to v', so the decay rate of the probability of his accepting this separation is given by $\varphi(v,v')$. m_2 could accept a proposed separation from w_2. This would change his payoff from u to u', so the decay rate of the probability of his accepting this separation is given by $\varphi(u,u')$. One of the women could accept a proposed separation from her partner. This would change her payoff from z to u', so the decay rate of the probability of her accepting this separation is given by $\varphi(z,u')$. Finally, one of the women could match with the man to whom she is not currently matched. This would increase the man's payoff, but would reduce the woman's payoff from z to u, so the decay rate of the probability of both parties accepting this rematching is given by $\varphi(z,u)$. Consequently, the most probable transition away from g_W has a decay rate equal to

$$\min\left\{\varphi(v,v'), \varphi(u,u'), \varphi(z,u'), \varphi(z,u)\right\}. \tag{4.1}$$

From g_M, there are the following types of transition to another matching. m_1 could accept a proposed separation from w_2. This would change his payoff from z to v', so the decay rate of the probability of his accepting this separation is given by $\varphi(z,v')$. m_2 could accept a proposed separation from w_1. This would change his payoff from z to u', so the decay rate of the probability of his accepting this separation is given by $\varphi(z,u')$. One of the women could accept a proposed separation from her partner. This would change her payoff from u to u', so the decay rate of the probability of her accepting this separation is given by $\varphi(u,u')$. m_1 and w_1 could agree to match. This would increase the payoff of w_1, but would reduce the payoff of m_1 from z to v, so the decay rate of the probability of both parties accepting this rematching is given by $\varphi(z,v)$. Finally, m_2 and w_2 could agree to match. This would increase the payoff of w_2, but would reduce the payoff of m_2 from z to u, so the decay rate of the probability of both parties accepting this rematching is given by $\varphi(z,u)$. Consequently, the most probable transition away from g_M has a decay rate equal to

$$\min\left\{\varphi(z,v'), \varphi(z,u'), \varphi(u,u'), \varphi(z,v), \varphi(z,u)\right\}. \tag{4.2}$$

Recall that $\varphi(v,v') \geq \varphi(u,u')$. It immediately follows that (4.2) is less than or equal to (4.1). That is, the most probable transition away from g_M is at least as probable as the most probable transition away from g_W.

Observe that in this example, following the initial transition away from a stable state, the unperturbed dynamic can reach the other stable state. Hence, by standard arguments (see, e.g. Young, 1993), $g_W \in SS$. Therefore $SS \nsubseteq Ra$.

	w_1	w_2	w_3
m_1	10, 1	5, 5	1, 10
m_2	1, 10	10, 1	5, 5
m_3	5, 5	1, 10	10, 1

(i) $\forall i \in N$, $u_i(\emptyset) = 0$.

	w_1	w_2	w_3
m_1	10, 4	5, 5	4, 10
m_2	4, 10	10, 4	5, 5
m_3	5, 5	4, 10	10, 4

(ii) $\forall i \in N$, $u_i(\emptyset) = 0$.

	w_1	w_2	w_3
m_1	0, 1	−5, 5	−9, 10
m_2	1, 10	10, 1	5, 5
m_3	5, 5	1, 10	10, 1

(iii) $u_{m_1}(\emptyset) = -10$ and $\forall i \in N \setminus \{m_1\}$, $u_i(\emptyset) = 0$.

Fig. 3. Payoffs discussed in the example of Section 5. Payoffs in **Panel (i)** are the payoffs of Example 4.1 of Newton and Sawa (2015). Payoffs in **Panel (ii)** are the payoffs in Panel (i) following an order-preserving transformation. Payoffs in **Panel (iii)** are the same as in Panel (i) except that payoffs of m_1 have been reduced by 10.

Step 2: Extend the argument to account for an arbitrary number of players.

When there are players other than m_1, m_2, w_1, w_2, we set utilities so that every man in M prefers every woman in $W \setminus \{w_1, w_2\}$ to every woman in $\{w_1, w_2\}$, and every woman in W prefers every man in $M \setminus \{m_1, m_2\}$ to every man in $\{m_1, m_2\}$. Further, we set utilities so that players in $N \setminus \{m_1, m_2, w_1, w_2\}$ would rather remain single than match with a player in $\{m_1, m_2, w_1, w_2\}$ and that at any stable matching, players in $N \setminus \{m_1, m_2, w_1, w_2\}$ match amongst themselves in a unique way. Consequently, there are two stable matchings, one at which players in $\{m_1, m_2, w_1, w_2\}$ match as they do at g_W, and one at which they match as they do at g_M. Denote these matchings $\tilde{g}_W$ and $\tilde{g}_M$. Further, let all payoffs for players in $N \setminus \{m_1, m_2, w_1, w_2\}$ be high enough that the Rawlsian stable matchings are determined by the payoffs of $\{m_1, m_2, w_1, w_2\}$, therefore $Ra = \{\tilde{g}_M\}$.

Note that if the process is at any matching at which the players in $\{m_1, m_2, w_1, w_2\}$ are not stably matched amongst themselves, then either of $\tilde{g}_W$ and $\tilde{g}_M$ can be reached by the unperturbed dynamic. Consider a path from $\tilde{g}_W$ to such a matching. The last step on this path must be the only step that involves one or two players in $\{m_1, m_2, w_1, w_2\}$. If it involves two such players, then the additional players are irrelevant and the argument of Step 1 can be applied. If it involves one such player and one player from outside of $\{m_1, m_2, w_1, w_2\}$, then our choice of utilities guarantees that it is only a deviation for the latter player. Consequently, its probability is independent of the payoff of the former player. Now, consider the rematchings on the path just considered, only this time starting from $\tilde{g}_M$. These rematchings are equally as probable as those on the earlier path and they terminate at a matching from which $\tilde{g}_W$ can be reached by the unperturbed dynamic. Consequently, once again we have $\tilde{g}_W \in SS$ and $SS \not\subseteq Ra$.

5. Comparison with other behavioral rules

Let $M = \{m_1, m_2, m_3\}$, $W = \{w_1, w_2, w_3\}$. We will consider the payoffs from each of the three panels of Fig. 3. Each player's ordinal payoff ranking is the same across panels, so the set of stable matchings remains the same. There are three stable matchings, which are

$$g_1 = \{m_1w_1, m_2w_2, m_3w_3\}, \qquad g_2 = \{m_1w_2, m_2w_3, m_3w_1\}, \qquad g_3 = \{m_1w_3, m_2w_1, m_3w_2\}.$$

Note that g_1 is man-optimal and g_3 is woman-optimal. We shall consider three different processes of strategy updating and see how moving across the different payoff specifications in Fig. 3 changes equilibrium selection in each case.

5.1. Uniform deviations

Consider best response with uniform deviations as defined in expression (3.5) in Section 3. Under this behavioral rule, all stable matchings are stochastically stable (Jackson and Watts, 2002). Therefore, g_1, g_2 and g_3 are all stochastically stable.

5.2. Logit choice

Now consider logit choice as defined in expression (3.6) in Section 3. Deviation probabilities under logit choice depend only on cardinal payoff losses for the players making the deviation. Under this behavioral rule, stochastically stable states will be among those from which the payoff loss required for a deviation to take place is as great as possible (Newton and Sawa, 2015). First consider the payoffs in Panel (i). From g_1, the lowest payoff loss from a deviation involves a woman becoming single and losing payoff of $1 - 0 = 1$. From g_3, the lowest payoff loss from a deviation involves a man becoming single and losing payoff of $1 - 0 = 1$. However, from g_2 the lowest payoff loss from a deviation involves a player leaving his current partner for his least preferred partner and losing payoff of $5 - 1 = 4$. Consequently, g_2 is uniquely stochastically stable. Adding or subtracting a constant from a player's payoff, as is done to the payoffs of m_1 in Panel (iii), does not change logit choice probabilities and hence g_2 remains uniquely stochastically stable. However, if payoffs undergo a transformation that preserves payoff order but does not preserve cardinal differences, then logit choice probabilities will change. For the

payoffs in Panel (ii), the lowest payoff loss from a deviation away from g_2 is $5-4=1$, whereas the lowest payoff loss from a deviation away from g_1 or g_3 is $4-0=4$. This, together with the symmetry of the problem, implies that the stochastically stable matchings are g_1 and g_3.

5.3. Condition dependence

Finally, consider condition dependent behavior. Theorem 1 of the current paper implies that stochastically stable matchings are contained in *Ra*. For the payoffs in Panel (i), $Ra=\{g_2\}$ and hence g_2 is uniquely stochastically stable. If all payoffs undergo an order-preserving transformation as in Panel (ii), this does not change *Ra* and g_2 remains uniquely stochastically stable. However, adding or subtracting a constant from a player's payoff, as is done to the payoffs of m_1 in Panel (iii), even though it leaves all players' individual orderings unchanged, may change interpersonal comparisons between players and hence change *Ra*. For the payoffs in Panel (iii), $Ra=\{g_1\}$ and hence g_1 is uniquely stochastically stable. This is the stable matching that gives m_1 the highest payoff. To see that this observation generalizes, note that if we reduce player i's payoffs enough, while keeping other players' payoffs fixed, then the Rawlsian stable matchings will be the stable matchings that maximize player i's payoff.

6. Conclusion

In this paper we have given a behavioral axiomatization of a social choice function. In marriage problems under perturbed pairwise best response dynamics, Rawlsian stable matchings emerge as long run social norms if and only if players follow condition dependent behavioral rules. There are a few possible directions for future research. First, there is the possibility that our result might extend to the roommate problem (one-sided matching) or the assignment problem (matching with transferable utility). Second, our approach could be followed to obtain behavioral axiomatizations of other social choice functions. Finally, similar to the approach pioneered by Sandholm (2005) for population games with negative externalities, one may study whether under a given class of behavioral rules, institutional aspects of the marriage market can be designed to facilitate the emergence of desired social norms. Such aspects might include marriage costs, divorce costs and restrictions on admissible partnerships due to geography and social status.

Appendix A. Additional notation and ancillary results

A.1. Costs of transitions

The identity of stochastically stable states depends on the transition probabilities of the process. To measure the limiting relative magnitude of these probabilities, a cost function is defined as follows.

Definition 5. The 1-step cost of the process moving from g to g' is defined as:

$$c(g,g') := \lim_{\eta\to 0} -\eta \log P_\eta(g,g'), \tag{A.1}$$

adopting the convention that $-\log 0=\infty$.

$c(g,g')$ is the exponential decay rate of the transition probability from g to g'. The rarer a transition, the higher its cost. Impossible transitions have infinite cost. Note that for $g\notin S$, there is a zero cost transition from g. This is because there is some $g'\neq g$, such that $P_\eta(g,g')$ does not approach zero as $\eta\to 0$.

For $g\neq g'$, recall that

$$P_\eta(g,g') = \begin{cases} F_g(i,j)\, A^i_\eta(g,g')\, A^j_\eta(g,g') & \text{if } g' = g+ij \text{ for some } ij\notin g,\\ F_g(i,j)\left(1-\left(1-A^i_\eta(g,g')\right)\left(1-A^j_\eta(g,g')\right)\right) & \text{if } g' = g-ij \text{ for some } ij\in g,\\ 0 & \text{if } g'\notin\{g+ij, g-ij\} \text{ for any } i\in M, j\in W, \end{cases} \tag{A.2}$$

and $P_\eta(g,g)=1-\sum_{g'\neq g} P_\eta(g,g')$.

The description of P_η in terms of A^k_η together with (3.4) allows $c(\cdot,\cdot)$ to be written as follows. For $g,g'\in G$, $g\neq g'$,

$$c(g,g') = \begin{cases} \varphi\big(u_i(g),u_i(g')\big)+\varphi\big(u_j(g),u_j(g')\big) & \text{if } g' = g+ij \text{ for some } ij\notin g,\\ \min\left\{\varphi\big(u_i(g),u_i(g')\big),\varphi\big(u_j(g),u_j(g')\big)\right\} & \text{if } g' = g-ij \text{ for some } ij\in g,\\ \infty & \text{otherwise.} \end{cases} \tag{A.3}$$

To prove Theorem 2 in Appendix C we also require notation for the overall cost of moving between g and g', even if many steps are required. We denote with $S(g,g')$ the set containing all sequences of matchings $(g_z)_{z=0}^k$, with $g_0=g$, $g_k=g'$, and $k\geq 1$.

Definition 6. The minimum cost of the process moving from g to g' is defined as:

$$C(g, g') := \min_{(g_z)_{z=0}^{k} \in S(g,g')} \sum_{z=0}^{k-1} c(g_z, g_{z'}). \tag{A.4}$$

A.2. The one-shot deviation principle

We call a transition $g \to g'$ from a matching $g \in G$ the *least cost deviation* from g if it has the lowest cost of all possible 1-step transitions from g.

Definition 7. Denote the set of possible least cost deviations from $g \in G$ by:

$$\mathrm{L}(g) := \arg\min_{g' \neq g} c(g, g')$$

and the set of pairs of players involved in least cost deviations from $g \in G$ by:

$$N_{\mathrm{L}}(g) := \{(i, j) \in M \times W : \exists\, g' \in \mathrm{L}(g) :\ g' = g - ij \text{ or } g' = g + ij\}$$

$c_{\mathrm{L}}(g)$ will be used to denote the cost of the least cost deviation from g.[8]

$$c_{\mathrm{L}}(g) := \min_{g' \neq g} c(g, g').$$

We use the word *deviation* as we shall be interested in the application of these concepts to $g \in S$.

Define OS, the set of matchings which are most robust to one-shot deviation:

$$OS = \left\{ g \in G :\ c_{\mathrm{L}}(g) = \max_{g' \in G} c_{\mathrm{L}}(g') \right\}.$$

As $c_{\mathrm{L}}(g)$ is strictly positive only for $g \in S$, it must be that $OS \subseteq S$. Newton and Sawa (2015) show that OS contains SS: a stochastically stable matching must be comparatively robust against one-shot deviation. If OS contains only one matching, then that matching must be uniquely stochastically stable.

Theorem 3 *(Newton and Sawa (2015), Theorem 3.4).* $SS \subseteq OS$.

Appendix B. Sufficiency of condition dependence for Rawlsian selection

Define MM, the set of stable matchings at which the players with the lowest payoff are matched players:

$$MM = \left\{ g \in S :\ \forall i \in \arg\min_{k} u_k(g),\ g(i) \neq \emptyset. \right\}$$

Lemma 1. *If $MM \neq S$, then $MM \cap OS = \emptyset$.*

Proof. Note that the set of unmatched players is the same at every $g \in S$. If no player is unmatched at any $g \in S$, then by definition of MM, we have that $MM = S$, which would contradict the assumption of the lemma that $MM \neq S$. Therefore, there must be a nonempty set of players that is unmatched at every $g \in S$. There are two cases to consider.

Case 1. $MM = \emptyset$.
If $MM = \emptyset$, then $MM \cap OS = \emptyset$ is immediate.

Case 2. $MM \neq \emptyset$.
Consider $g \in MM$. Let

$$j \in \arg\min_{k : g(k) = \emptyset} u_k(g). \tag{B.1}$$

By definition of MM, there exists i such that $g(i) \neq \emptyset$ and $u_i(g) < u_j(g)$. By (A.3),

[8] This differs from the concept of the radius of a stable state $g \in S$ (Ellison, 2000, citing a no longer extant working paper of Evans, 1993). The radius is defined as $R(g) = \min_{g' \in S \setminus \{g\}} C(g, g')$ and requires a different stable state to be reached by the process. It turns out that in the problems considered in the current paper $c_{\mathrm{L}}(g) = R(g)$ for all stable matchings outside of a specific set, but this does not follow from the definitions.

$$c(g, g - i\,g(i)) = \min\left\{\varphi\big(u_i(g), u_i(g - ig(i))\big), \varphi\big(u_{g(i)}(g), u_{g(i)}(g - ig(i))\big)\right\} \tag{B.2}$$
$$\leq \varphi\big(u_i(g), u_i(g - ig(i))\big).$$

Note that $g \in MM \subseteq S$ implies that $u_i(g) > u_i(g - ig(i))$. Let δ be such that $u_j(g) > \delta > u_i(g)$. We have,

$$c_L(g) \underbrace{\leq}_{\text{by (B.2)}} \varphi\big(u_i(g), u_i(g - ig(i))\big) \underbrace{<}_{\substack{\text{by condition dependence} \\ \text{and } \delta > u_i(g) > u_i(g - ig(i))}} \varphi\big(\delta, u_i(g - ig(i))\big) \tag{B.3}$$
$$\underbrace{\leq}_{\substack{\text{by condition dependence} \\ \text{and } u_j(g) > \delta > u_i(g - ig(i))}} \inf_{u' < u_j(g)} \varphi\big(u_j(g), u'\big).$$

Consider $g' \in S \setminus MM$. As the sets of unmatched players at g and g' are identical, we argue that every player at g' obtains a payoff of at least $u_j(g') = u_j(g)$. For unmatched players, this follows directly from (B.1). For matched players, it follows from $g' \notin MM$ and the definition of MM. Hence,

$$\min_k u_k(g') = u_j(g') = u_j(g). \tag{B.4}$$

As $g' \in S$, a transition from g' must involve at least one player losing payoff. Hence,

$$c_L(g') \underbrace{\geq}_{\text{by (A.3)}} \min_k \inf_{u' < u_k(g')} \varphi\big(u_k(g'), u'\big) \tag{B.5}$$
$$\underbrace{=}_{\substack{\text{as condition dependence implies} \\ \text{that lowering } u_k \text{ gives lower } \varphi(u_k, u') \\ \text{for all } u' < u_k, \text{ hence minimum} \\ \text{attained when } u_k \text{ is minimized}}} \inf_{u' < \min_k u_k(g')} \varphi\left(\min_k u_k(g'), u'\right) \underbrace{=}_{\text{by (B.4)}} \inf_{u' < u_j(g)} \varphi\big(u_j(g), u'\big).$$

Expressions (B.3) and (B.5), taken together, imply $c_L(g) < c_L(g')$. It follows from the definition of OS that $g \notin OS$.

As the above argument holds for any $g \in MM$, it follows that $MM \cap OS = \emptyset$. *Q.E.D.*

Lemma 2. *If $MM \neq S$, then $Ra = S \setminus MM$.*

Proof. Note that the set of unmatched players is the same at every $g \in S$. If no player is unmatched at any $g \in S$, then by definition of MM, we have that $MM = S$, which would contradict the assumption of the lemma that $MM \neq S$. Therefore, there must be a nonempty set of players that is unmatched at every $g \in S$.

For some $g \in S$, define j as in expression (B.1). Note that as the set of unmatched players and hence the payoffs of unmatched players are the same at every stable matching, it does not matter which $g \in S$ we use. Let u_j^* be the payoff of j at every stable matching. It follows from the definition of MM that

$$\min_k u_k(g) < u_j^* \quad \text{for all} \quad g \in MM, \tag{B.6}$$

and

$$\min_k u_k(g) = u_j^* \quad \text{for all} \quad g \in S \setminus MM. \tag{B.7}$$

From (B.6), (B.7) and the definition of Ra, it follows that if $g \in MM$, then $g \notin Ra$, and if $g \in S \setminus MM$, then $g \in Ra$. *Q.E.D.*

Lemma 3. *If $MM \neq S$, then $OS \subseteq Ra$.*

Proof.

$$OS \underbrace{\subseteq}_{\text{by Lemma 1}} S \setminus MM \underbrace{=}_{\text{by Lemma 2}} Ra. \quad \square \tag{B.8}$$

Q.E.D.

Lemma 4. *If $MM = S$, then $OS \subseteq Ra$.*

Proof. Consider $g \in MM$ and $i^* \in \arg\min_k u_k(g)$.

By definition of MM, $g(i^*) \neq \emptyset$.

Note that $g \in MM \subseteq S$ implies that $u_{i^*}(g) > u_{i^*}(g - i^* g(i^*))$. Therefore,

$$c_L(g) \underbrace{\leq}_{\substack{\text{by defn of} \\ c_L(\cdot)}} c(g, g - i^* g(i^*)) \underbrace{\leq}_{\text{by (A.3)}} \varphi\big(u_{i^*}(g), u_{i^*}(g - i^* g(i^*))\big) \tag{B.9}$$

$$\underbrace{\leq}_{\substack{\text{by defn of sup} \\ \text{and } u_{i^*}(g) > u_{i^*}(g - i^* g(i^*))}} \sup_{u' < u_{i^*}(g)} \varphi\big(u_{i^*}(g), u'\big) \underbrace{=}_{\substack{\text{by defn} \\ \text{of } i^*}} \sup_{u' < \min_k u_k(g)} \varphi\left(\min_k u_k(g), u'\right).$$

As $g \in S$, a transition from g must involve at least one player losing payoff. Hence,

$$c_L(g) \underbrace{\geq}_{\text{by (A.3)}} \min_k \inf_{u' < u_k(g)} \varphi\big(u_k(g), u'\big). \tag{B.10}$$

Note that, by definition, $Ra \neq \emptyset$.

If $Ra = S$, then $OS \subseteq S = Ra$ and we are done.

If $Ra \neq S$, let $g' \in Ra$ and $g'' \in S \setminus Ra$.

As $MM = S$ (as assumed in the lemma statement), we have $g', g'' \in MM$ and therefore (B.9) and (B.10) apply to g' and g''.

Furthermore, by definition of Ra, we have that $\min_k u_k(g') > \min_k u_k(g'')$. Let δ', δ'' be such that $\min_k u_k(g') > \delta' > \delta'' > \min_k u_k(g'')$. Thus,

$$c_L(g') \underbrace{\geq}_{\text{by (B.10)}} \min_k \inf_{u' < u_k(g')} \varphi\big(u_k(g'), u'\big) \underbrace{=}_{\substack{\text{as condition dependence implies that lowering } u_k \text{ gives lower } \varphi(u_k, u') \\ \text{for all } u' < u_k, \text{ hence minimum} \\ \text{attained when } u_k \text{ is minimized}}} \inf_{u' < \min_k u_k(g')} \varphi\left(\min_k u_k(g'), u'\right) \tag{B.11}$$

$$\underbrace{\geq}_{\substack{\text{by condition dependence} \\ \text{and } \min_k u_k(g') > \delta' > \min_k u_k(g'')}} \varphi\left(\delta', \min_k u_k(g'')\right) \underbrace{>}_{\substack{\text{by condition dependence} \\ \text{and } \delta' > \delta'' > \min_k u_k(g'')}} \varphi\left(\delta'', \min_k u_k(g'')\right)$$

$$\underbrace{\geq}_{\substack{\text{by condition dependence} \\ \text{and } \delta'' > \min_k u_k(g'')}} \sup_{u' < \min_k u_k(g'')} \varphi\big(\min_k u_k(g''), u'\big) \underbrace{\geq}_{\text{by (B.9)}} c_L(g'')$$

Expression (B.11) gives $c_L(g') > c_L(g'')$. It follows from the definition of OS that $g'' \notin OS$.

As the above argument holds for any $g'' \notin Ra$, it follows that $OS \subseteq Ra$. *Q.E.D.*

Proof of Theorem 1.

$$SS \underbrace{\subseteq}_{\text{by Theorem 3}} OS \underbrace{\subseteq}_{\text{by Lemmas 3, 4}} Ra. \quad \textit{Q.E.D.} \tag{B.12}$$

Appendix C. Necessity of condition dependence for Rawlsian selection

Proof of Theorem 2. In the following we show by construction that $\exists u \in \mathcal{U}$ such that: $S = \{g, g'\}$, $g \notin Ra$, and $g \in SS$.

Suppose w.l.o.g. that $l = |W| \leq |M| = k$. We set $g = \{m_1 w_1, m_2 w_2\} \cup \overline{g}$, $g' = \{m_1 w_2, m_2 w_1\} \cup \overline{g}$, with $\overline{g} = \{m_{i+k-l} w_i | \forall i \in \mathbb{N}, 2 < i \leq l\}$. Preferences of players in $\{m_1, m_2, w_1, w_2\}$ over themselves are denoted as in Table 1, where $z > u > v > v'$, $u > u'$; moreover, we set $u_i(\{ij\}) > z\ \forall (i, j)$ such that either $i \in \{m_1, m_2\}$ and $j \in W \setminus \{w_1, w_2\}$, or $i \in \{w_1, w_2\}$ and $j \in M \setminus \{m_1, m_2\}$. Finally, for $i \in M \setminus \{m_1, m_2\}$, we set $u_i(\{iw_r\}) > u_i(\{iw_{r'}\})$ whenever $r > r'$, and $u_i(\{iw_3\}) > u_i(\emptyset) > z > u_i(\{iw_2\})$. Analogously, for $i \in W \setminus \{w_1, w_2\}$, we set $u_i(\{im_r\}) > u_i(\{im_{r'}\})$ whenever $r > r'$, and $u_i(\{im_3\}) > u_i(\emptyset) > z > u_i(\{im_2\})$. Given the listed inequalities on preferences, $S = \{g, g'\}$, and $g \notin Ra$.

In order to determine which matchings are stochastically stable, we start by identifying $C(g, g')$. Since any single addition/deletion of an edge involving at least one player in $\{m_1, m_2, w_1, w_2\}$ is enough to reach a matching from which g' is reachable in the unperturbed dynamics (hence, $C(\hat{g}, g') = 0$), and given that $C(g, g')$ is the minimum cost of the process moving from g to g', we focus on sequences where all initial steps do not involve any player in $\{m_1, m_2, w_1, w_2\}$, followed

Table 1
Payoffs for players in $\{m_1, m_2, w_1, w_2\}$. The letter in each cell of the matrix refers to the payoff earned by the row player when matched to the column player. When a player is matched to him or herself, it is to be intended as remaining alone.

player	m_1	m_2	w_1	w_2
m_1	v'	/	v	z
m_2	/	u'	z	u
w_1	z	u	u'	/
w_2	u	z	/	u'

Table 2
Cost of the single step involving at least one agent in $\{m_1, m_2, w_1, w_2\}$, when moving from g to g' and from g' to g.

$c(\tilde{g}, \hat{g})$	step	$c(\tilde{g}', \hat{g}')$	step
$\min\{\varphi(v, v'), \varphi(z, u')\}$	$-m_1w_1$	$\min\{\varphi(z, v'), \varphi(u, u')\}$	$-m_1w_2$
$\min\{\varphi(u, u'), \varphi(z, u')\}$	$-m_2w_2$	$\min\{\varphi(z, u'), \varphi(u, u')\}$	$-m_2w_1$
$\varphi(z, u)$	$+m_1w_2$	$\varphi(z, v)$	$+m_1w_1$
$\varphi(z, u)$	$+m_2w_1$	$\varphi(z, u)$	$+m_2w_2$
$\varphi(u_{\tilde{i}}(\tilde{g}), u_{\tilde{i}}(\tilde{g} + \tilde{i}\tilde{j}))$	$+\tilde{i}\tilde{j}$	$\varphi(u_{\tilde{i}'}(\tilde{g}'), u_{\tilde{i}}(\tilde{g}' + \tilde{i}'\tilde{j}'))$	$+\tilde{i}'\tilde{j}'$

by a step where at least one player in $\{m_1, m_2, w_1, w_2\}$ is involved, with all remaining steps occurring in the unperturbed dynamics.

Suppose that a sequence with minimum cost has initial steps that lead from $g = \{m_1w_1, m_2w_2\} \cup \overline{g}$ to $\tilde{g} := \{m_1w_1, m_2w_2\} \cup \overline{\overline{g}}$ (possibly, no such initial steps are needed and $\overline{g} = \overline{\overline{g}}$),[9] and then a step from $\tilde{g}$ to $\hat{g} := \{m_1w_1, m_2w_2\} \cup \overline{\overline{g}} \pm ij$, with $\{i, j\} \cap \{m_1, m_2, w_1, w_2\} \neq \emptyset$. We can hence write

$$C(g, g') = C(g, \tilde{g}) + c(\tilde{g}, \hat{g}) + C(\hat{g}, g'), \tag{C.1}$$

with

$$C(g, \tilde{g}) = C\left(\{m_1w_1, m_2w_2\} \cup \overline{g}, \{m_1w_1, m_2w_2\} \cup \overline{\overline{g}}\right), \tag{C.2}$$

$$c(\tilde{g}, \hat{g}) = c\left(\{m_1w_1, m_2w_2\} \cup \overline{\overline{g}}, \{m_1w_1, m_2w_2\} \cup \overline{\overline{g}} + ij\right), \tag{C.3}$$

$$C(\hat{g}, g') = 0. \tag{C.4}$$

Next, we show that there exists a sequence from g' to g with an associated cost that is smaller than or equal to $C(g, g')$. In particular, we consider

$$C(g', \tilde{g}') + c(\tilde{g}', \hat{g}') + C(\hat{g}', g) \tag{C.5}$$

where $\tilde{g}' := \{m_1w_2, m_2w_1\} \cup \overline{\overline{g}}$, and $\hat{g}' := \tilde{g}' \pm i'j'$, for some $i'j'$ such that $\{i', j'\} \cap \{m_1, m_2, w_1, w_2\} \neq \emptyset$. We note that from $\tilde{g}'$ the process can move to g in the unperturbed dynamics, hence $C(\hat{g}', g) = 0$. Also, the process can move from g' to $\tilde{g}$ with cost

$$C\left(\{m_1w_2, m_2w_1\} \cup \overline{g}, \{m_1w_2, m_2w_1\} \cup \overline{\overline{g}}\right), \tag{C.6}$$

which is equal to the quantity in (C.2), because no transitions on the relevant paths for (C.2) and (C.6) involve any player in $\{m_1, m_2, w_1, w_2\}$. We are left to show that $c(\tilde{g}', \hat{g}') \leq c(\tilde{g}, \hat{g})$, of which all possible values are reported in Table 2, where $\tilde{i} \notin \{m_1, m_2, w_1, w_2\}$, $\tilde{j} \in \{m_1, m_2, w_1, w_2\}$. We observe that any cost in the list from $\tilde{g}$ to $\hat{g}$ appears in the list from $\tilde{g}'$ to $\hat{g}'$ as well, but for $\varphi(v, v')$. However, we know that $\varphi(v, v') \geq \varphi(u, u')$ (the latter being in the list from $\tilde{g}'$ to $\hat{g}'$).

Therefore, we can conclude that $C(g, g') \geq C(g', \tilde{g}') + c(\tilde{g}', \hat{g}') + C(\hat{g}', g) \geq C(g', g)$, which in turn implies $g \in SS$. *Q.E.D.*

References

Agrawal, A., 2002. Genetic loads under fitness-dependent mutation rates. J. Evol. Biol. 15, 1004–1010.
Alós-Ferrer, C., Netzer, N., 2010. The logit-response dynamics. Games Econ. Behav. 68, 413–427.
Bergin, J., Lipman, B.L., 1996. Evolution with state-dependent mutations. Econometrica 64, 943–956.

[9] Actually, initial steps are never present in a sequence with minimum cost when the following single step involves two players in $\{m_1, m_2, w_1, w_2\}$.

Bilancini, E., Boncinelli, L., 2020. The evolution of conventions under condition-dependent mistakes. Econ. Theory 69, 497–521.
Bilancini, E., Boncinelli, L., Nax, H., 2020. What noise matters? Experimental evidence for stochastic deviations in social norms. Mimeo.
Biró, P., Bomhoff, M., Golovach, P.A., Kern, W., Paulusma, D., 2013. Solutions for the stable roommates problem with payments. In: Graph-Theoretic Concepts in Computer Science. In: Lecture Notes in Computer Science. Springer Verlag, Berlin, Germany, pp. 69–80.
Blume, L.E., 1993. The statistical mechanics of strategic interaction. Games Econ. Behav. 5, 387–424.
Blume, L.E., 2003. How noise matters. Games Econ. Behav. 44, 251–271.
Chen, B., Fujishige, S., Yang, Z., 2016. Random decentralized market processes for stable job matchings with competitive salaries. J. Econ. Theory 165, 25–36.
Cournot, A.A., 1838. Recherches sur les principes mathématiques de la théorie des richesses par Augustin Cournot. Chez L. Hachette.
Doğan, B., Klaus, B., 2018. Object allocation via immediate-acceptance: characterizations and an affirmative action application. J. Math. Econ. 79, 140–156.
Ehlers, L., Klaus, B., 2016. Object allocation via deferred-acceptance: strategy-proofness and comparative statics. Games Econ. Behav. 97, 128–146.
Ellison, G., 2000. Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution. Rev. Econ. Stud. 67, 17–45.
Foster, D., Young, H.P., 1990. Stochastic evolutionary game dynamics. Theor. Popul. Biol. 38, 219–232.
Gale, D., Shapley, L.S., 1962. College admissions and the stability of marriage. Am. Math. Mon. 69, 9–15.
Gillies, D.B., 1959. Solutions to general non-zero-sum games. Contrib. Theory Games 4, 47–85.
Hwang, S.H., Lim, W., Neary, P., Newton, J., 2018a. Conventional contracts, intentional behavior and logit choice: equality without symmetry. Games Econ. Behav. 110, 273–294.
Hwang, S.H., Lim, W., Neary, P., Newton, J., 2018b. Conventional contracts, intentional behavior and logit choice: equality without symmetry. Games Econ. Behav. 110, 273–294.
Jackson, M.O., Watts, A., 2002. The evolution of social and economic networks. J. Econ. Theory 106, 265–295.
Kalai, E., 1977. Proportional solutions to bargaining situations: interpersonal utility comparisons. Econometrica 45, 1623–1630.
Kalai, E., Smorodinsky, M., 1975. Other solutions to Nash's bargaining problem. Econometrica 43, 513–518.
Kandori, M., Mailath, G.J., Rob, R., 1993. Learning, mutation, and long run equilibria in games. Econometrica 61, 29–56.
Klaus, B., Newton, J., 2016. Stochastic stability in assignment problems. J. Math. Econ. 62, 62–74.
Klaus, B., Payot, F., 2015. Paths to stability in the assignment problem. J. Dyn. Games 2, 257–287.
Kojima, F., Manea, M., 2010. Axioms for deferred acceptance. Econometrica 78, 633–653.
Kojima, F., Ünver, M.U., 2014. The "Boston" school-choice mechanism: an axiomatic approach. Econ. Theory 55, 515–544.
Lim, W., Neary, P.R., 2016. An experimental investigation of stochastic adjustment dynamics. Games Econ. Behav. 100, 208–219.
Ma, J., 1994. Strategy-proofness and the strict core in a market with indivisibilities. Int. J. Game Theory 23, 75–83.
Maruta, T., 2002. Binary games with state dependent stochastic choice. J. Econ. Theory 103, 351–376.
Mäs, M., Nax, H.H., 2016. A behavioral study of "noise" in coordination games. J. Econ. Theory 162, 195–208.
Masarani, F., Gokturk, S.S., 1989. On the existence of fair matching algorithms. Theory Decis. 26, 305–322.
Maschler, M., Peleg, B., Shapley, L.S., 1979. Geometric properties of the kernel, nucleolus, and related solution concepts. Math. Oper. Res. 4, 303–338.
Nash, John F.J., 1950. The bargaining problem. Econometrica 18, 155–162.
Nax, H.H., Pradelski, B.S.R., 2015. Evolutionary dynamics and equitable core selection in assignment games. Int. J. Game Theory 44, 903–932.
Nax, H.H., Pradelski, B.S.R., Young, H.P., 2013. Decentralized dynamics to optimal and stable states in the assignment game. In: Proceedings of the 52nd IEEE Conference on Decision and Control, pp. 2391–2397.
Newton, J., 2018. Evolutionary game theory: a renaissance. Games 9, 31.
Newton, J., 2019. Conventions under heterogeneous choice rules. SSRN Working Paper Series 3383471.
Newton, J., Sawa, R., 2015. A one-shot deviation principle for stability in matching problems. J. Econ. Theory 157, 1–27.
Peski, M., 2010. Generalized risk-dominance and asymmetric dynamics. J. Econ. Theory 145, 216–248.
Rawls, J., 1971. A Theory of Justice. Harvard University Press.
Roth, A.E., Vande Vate, J.H., 1990. Random paths to stability in two-sided matching. Econometrica 58, 1475–1480.
Sandholm, W.H., 2005. Negative externalities and evolutionary implementation. Rev. Econ. Stud. 72, 885–915.
Sawa, R., 2014. Coalitional stochastic stability in games, networks and markets. Games Econ. Behav. 88, 90–111.
Svensson, L.G., 1999. Strategy-proof allocation of indivisible goods. Soc. Choice Welf. 16, 557–567.
Young, H.P., 1993. The evolution of conventions. Econometrica 61, 57–84.