Unilateral Stability in the Roommate Problem*

Michael Richter

Baruch College, City University of New York
and Royal Holloway, University of London

Ariel Rubinstein

School of Economics, Tel Aviv University
and Department of Economics, New York University

August 12, 2023

ABSTRACT: The roommate problem is commonly studied under the premise that harmony is disrupted by the intentional establishment of a new match. We instead focus on scenarios in which harmony is disrupted by a single agent who unilaterally initiates contact with a member of a different pair, regardless of whether or not his approach is reciprocated. A variety of solution concepts are proposed in which taboos, status, or power systematically limit such initiatives in order to achieve harmony.

KEYWORDS: Roommate problem, pairwise stability, unilateral stability.
AEA Classification: C78.

*The Orchard Tea Garden in Cambridge was the place where we were inspired to write this paper. We are grateful to Áron Tóbiás for his very useful comments.

1. Introduction

Matching problems form a class of iconic models in economic theory beginning with Gale and Shapley (1962). Classically, the threat to stability in those models is two members agreeing to form a new match. In contrast, we focus on scenarios in which instability arises from a single agent unilaterally approaching another in a different pair, which is akin to the stability threat underlying competitive equilibria or Nash equilibria.

As a platform we use the roommate problem, a closely related variant of the two-sided matching problem. In the model, N is a single even-numbered population of n agents who must partition themselves into pairs. Each agent i is characterized by his strict preferences $\succ^i$ over the other agents. A *pairing* is a profile $(x^i)_{i\in N}$ that specifies for every i a partner $x^i \neq i$, such that for any two agents i and j, if $x^i = j$ then $x^j = i$. To denote a match between i and j, we write $i \leftrightarrow j$. A pairing is said to be *pairwise stable* if there are no two agents in different pairs, each of whom prefers the other to his current partner. Often a pairwise stable pairing does not exist. The familiar modifications of pairwise stability preserve the voluntary formation of a new match as the basic threat to stability (see for example Morrill (2010),[1] Root and Ahn (2020),[2] and Tan (1990)[3]).

Contrary to previous work, our focus in this paper is not on the non-existence of pairwise stable pairings. Rather, we are motivated by the desire to model situations where the overall harmony of the system could be jeopardized by a unilateral move by a single agent.

Here are two leading scenarios we have in mind:

Scenario I: All agents are paired up. Any agent is able to initiate contact with any other and propose that they form a new match. Social harmony is disturbed not only by the actual formation of this proposed match but even by the mere act of one agent approaching another with such a proposal. The result is instability since once an ap-

[1] Morrill (2010) takes the roommate problem literally, stating "stability ignores the key physical constraint that roommates require a room and is therefore too restrictive." This leads him to propose an efficient algorithm for finding a Pareto improvement subject to an initial assignment.

[2] Root and Ahn (2020) find that a generalized serial dictatorship procedure is the only mechanism that satisfies Pareto-efficiency and group strategy-proofness.

[3] Tan (1990) proposes and analyzes a second-best stability concept: "Since there may not exist any complete stable matching, it is natural to consider the problem of finding a maximum number of disjoint pairs of persons such that these pairs are stable among themselves, i.e. no two persons, who are not paired together but have matched partners, both prefer each other to their partners under the matching."

proach is made, a scandal bursts: the betrayed partner feels resentment, and regardless of whether the approach is reciprocated, bad sentiments spread throughout society.

In other words, a society can be destabilized by an agent A approaching B and expressing his desire for B to abandon his current partner and match with him instead. There are several possible motives for A's approach: He might know that B also prefers him over B's current partner (which is the premise of pairwise stability). Alternatively, even if he knows that B does not prefer him, A might hope that if he approaches B, then B will feel flattered and change his opinion. Finally, A might lack knowledge about B's preferences and simply tries his luck.

Scenario II: In an organization, each office is shared by two employees. Every employee has access to the box of office keys and can take any key he wishes. In such a situation, the source of instability is not the establishment of a blocking pair by mutual agreement, but rather the unilateral move of one employee taking someone else's key and thus forcing the formation of a new partnership.

Except for very rare preference configurations, every agent cannot be paired with his top choice, and therefore, if no restrictions are placed on who can approach whom (or who can take whose keys), then there is no unilaterally stable pairing of the agents. In this paper, we take the approach that stability in such situations can be restored by restricting agents' ability to approach other agents (in Scenario I) or to take the keys of others (in Scenario II). In Scenario I, a familiar conservative norm forbids approaching any matched individual. Such a norm solves all instability problems, but at the cost of dramatically restricting individual liberty. The analogue of this norm in Scenario II is an organizational rule prohibiting agents from changing offices. We instead examine less restrictive norms that echo familiar social institutions: **taboos**, **status** and **power**.

We propose and analyze several solution concepts which employ taboos, status or power to restrict unilateral moves in order to obtain harmony. The first is T-equilibrium, which models an adherence to social norms regarding which matches are permissible and which are taboo. The next two equilibrium notions are inspired by competitive equilibrium concepts, except that a status ranking prevails instead of prices. In the first of the two, called C-equilibrium, an agent can only approach agents who are lower-ranked than *himself*. In the other, called S-equilibrium, an agent can only approach

agents who are lower-ranked than his *partner*. Finally, in our J-equilibrium variants, there is a power hierarchy which determines who can approach whom. As a mnemonic device, we associate the letter T with taboo, S with status, C with competitive, and J with jungle.

Let us emphasize again that we do not deny that pairwise stability is often a natural solution concept. However, we do not believe that it should be the only criterion for harmony that might be of interest in matching models.

2. The permissible and the forbidden: T-equilibrium

The first solution concept is inspired by Richter and Rubinstein (2020). This equilibrium reflects a social norm that determines which couples are permissible and which are forbidden. An equilibrium is comprised of a pairing and a permissible set of matches such that: (i) each agent's partner is his most preferred from among those with whom he can form a permissible pair, and (ii) the set of permissible matches is maximal in the sense that for any strict superset of permissible couples there is no pairing for which (i) is true.

As discussed in the introduction, if everything is permissible, then harmony is typically impossible. On the flip side, if everything (besides the current pairing) is forbidden, then harmony can be achieved, but at a terrible cost to freedom. The T-equilibrium concept strikes a balance between freedom and harmony. It allows as much freedom as possible without compromising the achievement of harmony.

Given a set of permissible matches Y, agent i's choice set is $C_i(Y)=\{j \mid i \leftrightarrow j \in Y\}$, which is the set of partners with whom his match is permitted. Notice that if i is permitted to choose j, then j is permitted to choose i. A *para-T-equilibrium* is a tuple $\langle Y,(x^i)\rangle$ where Y is a set of couples and (x^i) is a pairing such that for every agent i, agent x^i is i's most preferred in $C_i(Y)$. A *T-equilibrium* is a para-T-equilibrium such that there is no other para-T-equilibrium $\langle Z,(y^i)\rangle$ with a larger permissible set $Y \subset Z$.

Claim 1 characterizes the set of T-equilibrium pairings as the set of Pareto-optimal pairings (consequently, a T-equilibrium always exists).

Claim 1: A pairing is a T-equilibrium pairing if and only if it is Pareto-optimal.

Proof: Given a pairing (x^i), define $L((x^i))$ to be the set of all matches $i \leftrightarrow j$ such that $x^i \succsim^i j$ and $x^j \succsim^j i$. That is, it is the set of couples that are weakly less-desired by both of the involved agents relative to the pairing (x^i). Note that for every i the match $i \leftrightarrow x^i$ is in $L((x^i))$. Clearly, $\langle L((x^i)), (x^i)\rangle$ is a para-T-equilibrium since all pairs whose permissibility would disrupt the harmony of the pairing (x^i) are forbidden.

Let $\langle Y, (x^i)\rangle$ be a T-equilibrium. By the maximality of Y, it must be that $Y = L((x^i))$. If there exists a pairing (y^i) that Pareto-dominates (x^i), then the tuple $\langle L((y^i)), (y^i)\rangle$ is a para-T-equilibrium. Obviously, $L((y^i)) \supseteq L((x^i))$ and this inclusion is strict, since at least one agent, say j, is strictly better off in (y^i), which means that the match $j \leftrightarrow y^j$ is in $L((y^i))$ but not in $L((x^i))$. Having a para-T-equilibrium with a larger set of permissible pairs contradicts the supposition of $\langle Y, (x^i)\rangle$ being a T-equilibrium.

On the other hand, let (x^i) be a Pareto-optimal pairing. As previously mentioned, $\langle L((x^i)), (x^i)\rangle$ is a para-T-equilibrium. To demonstrate that it is a T-equilibrium, suppose to the contrary that there exists a para-T-equilibrium $\langle Y, (y^i)\rangle$ with $Y \supset L((x^i))$. In this scenario, all agents are weakly better off in (y^i) since for each agent i, agent x^i is available in Y. Moreover, the set Y contains at least one potential match $i \leftrightarrow j$ not in $L((x^i))$ for which $j \succ^i x^i$. However, in that case $y^i \succsim^i j \succ^i x^i$ and thus (y^i) Pareto-dominates (x^i), a contradiction. ■

Throughout the analysis, we will consider two illustrative examples. In the first, which is referred to as the *common-ranking roommate economy*, there is a common ranking $i_1 \succ i_2 \succ \cdots \succ i_n$ and all agents prefer each other according to this order (that is, for each i and every $j, k \neq i$, $j \succ^i k$ if and only if $j \succ k$). Thus, there is a total conflict of interests between the agents. In the second example, termed the *reciprocated-love roommate economy*, every agent's true love is reciprocated, that is, if i top-ranks j, then j top-ranks i. Thus, there is no conflict between agents' preferences.

Common Ranking: In the common-ranking roommate economy, every pairing is Pareto-optimal, and therefore is a T-equilibrium pairing. The unique T-equilibrium with the pairing $\{i_1 \leftrightarrow i_2, i_3 \leftrightarrow i_4, \ldots\}$ has a particularly simple structure: every agent has exactly one permitted match, and all other matches are forbidden.

Reciprocated Love: On the opposite end, in the reciprocated-love roommate economy where there is no conflict in preferences, there is a unique T-equilibrium. Its permissible set includes all possible matches, reflecting the harmonious alignment of preferences among the agents.

3. Status ranking

In this section, we examine scenarios in which harmony is achieved by means of a status ranking among the agents. In real life, such rankings are sometimes external and due to factors outside our domain, such as caste or physical appearance. Here we have in mind situations in which the ranking is endogenous, just as prices are in a competitive equilibrium. This internal ranking is responsive to the configuration of preferences, reflecting statements such as "the most popular individual is the highest ranked" and "the most popular individual determines the ranking of the others".

In the following two subsections, we examine two solution concepts that utilize status as a means to achieve harmony. In the first concept, an agent can only be matched with an agent of a weakly lower status than his own. In the second concept, an agent's "wealth" is determined in equilibrium by the status of his matched partner, and no agent can approach any agent who has a higher status than his partner's. For both concepts, in equilibrium, the agents' optimal choices given the status ranking result in a pairing.

In the spirit of competitive equilibrium, a status ranking can be thought of as a measure of value, and an agent chooses his optimal match given his own value (or that of his partner). In this interpretation, status imposes a physical restriction on an agent's choices. An alternative interpretation of the status ranking is that of anti-prestige, meaning that a higher rank corresponds to lower prestige (and a lower rank corresponds to higher prestige). Agents cannot bear any loss of prestige, and thus will only consider approaching agents with a weakly lower ranking, i.e. weakly more prestigious. In this interpretation, status distorts an agent's views of potential partners.

3.1 C-equilibrium

A *C-equilibrium* is a tuple $\langle \trianglerighteq, (x^i) \rangle$ where the statement $i \trianglerighteq j$ is interpreted as "i has weakly higher status than j". A social institution prevents an agent from being matched with anyone who is strictly higher ranked than *himself*. In a *C-equilibrium*, for every agent i, his partner x^i is i's most preferred roommate in his choice set $\{j \in N - \{i\} \mid i \trianglerighteq j\}$. Thus, the ordering $\trianglerighteq$ shapes the agents' "budget sets". This definition is related to a special case of the notion of UE^e (Richter and Rubinstein, 2015), in which each agent's initial endowment is himself. Note that in every C-equilibrium, the status of any two matched agents must be identical.

In order to characterize C-equilibria, we introduce the concept of *pair-rankability*. A roommate economy is said to be *pair-rankable* if there exists a partition of N into doubletons $\{I_1, \ldots, I_{n/2}\}$ with the property that, for every i and q, if $i \in I_q$, then i top-ranks his doubleton's partner from the union of the later doubletons $I_q \cup \cdots \cup I_{n/2}$.[4] The pairing generated by such a partition of N is clearly the unique pairwise stable pairing.

The following claim demonstrates that a C-equilibrium exists if and only if the economy is pair-rankable.

Claim 2: (i) An economy has a C-equilibrium if and only if it is pair-rankable.
(ii) If a C-equilibrium pairing exists, then it is unique and pairwise stable.

Proof: (i) Let $\langle \trianglerighteq, (x^i) \rangle$ be a C-equilibrium and let i be a $\trianglerighteq$-maximal agent. As mentioned earlier, i and x^i must be equally $\trianglerighteq$-ranked, and thus both agents must top-rank each other. These two agents form I_1 and are removed. This process repeats itself with the remaining agents to eventually form $I_2, \ldots, I_{n/2}$. Thus, the economy is pair-rankable.

Assume that the economy is pair-rankable with the sequence of doubletons $I_1, \ldots, I_{n/2}$. Define $i \trianglerighteq j$ if $i \in I_q$, $j \in I_r$ and $q \leq r$. Define $x^i = j$ if $\{i, j\}$ is one of the doubletons. Clearly, the tuple $\langle \trianglerighteq, (x^i) \rangle$ constitutes a C-equilibrium and its pairing is pairwise stable.

[4] Pair-rankability is a milder condition than α-reducibility (Alcalde, 1995), which is used to ensure the existence of a pairwise-stable pairing in the roommate problem. Two classic pair-rankable cases are:
(i) Agents residing in a metric space who prefer closer agents.
(ii) Agents positioned on a line with single-peaked preferences over other agents, where the peak is always one of their neighbors. This implies that an extreme agent top-ranks his only neighbor, and for any set of agents, there exist two neighbors such that the left one top-ranks his right neighbor and the right one top-ranks his left neighbor.

(ii) A proof by induction on the number of agents: Assume that a C-equilibrium exists. The status of any two matched agents must be the same. In particular, the highest-status agent is matched with another highest-status agent, and they both must top-rank each other (recall that preferences are strict). Consequently, they must be matched in any C-equilibrium (since for any equilibrium ranking, one of the two agents can "afford" the other and would prefer him over any other potential partner). Any C-equilibrium induces a C-equilibrium among the remaining agents. By the induction hypothesis, this induced C-equilibrium pairing is unique among the remaining agents. Therefore, the C-equilibrium pairing is unique, and by part (i) it is pairwise stable. ■

Common Ranking: The common ranking roommate economy is uniquely pair-rankable with $I_q = \{i_{2q-1}, i_{2q}\}$. Therefore, there is a unique C-equilibrium: the pairing is $\{i_1 \leftrightarrow i_2, i_3 \leftrightarrow i_4, \ldots\}$ and the status ranking is $i_1 \sim i_2 \rhd i_3 \sim i_4 \rhd \cdots \rhd i_{n-1} \sim i_n$.

Reciprocated Love: The C-equilibrium pairing is unique, each agent ends up with their soulmate; however, the status ranking is quite free and can even be total indifference.

3.2 S-equilibrium

An S-equilibrium candidate also consists of a pairing and a status ordering of the agents. But here, a social institution prevents an agent from approaching any agent who is ranked higher than *his partner* (rather than himself, as in a C-equilibrium). In equilibrium, no agent can find a different partner whom he judges to be more desirable and who has a weakly lower status than his current partner.

Formally, a tuple $\langle \trianglerighteq, (x^i) \rangle$ is an *S-equilibrium* if, for every agent i, there is no j such that $j \succ^i x^i$ and $x^i \trianglerighteq j$. In this context, we find the previously mentioned anti-prestige interpretation particularly compelling: an agent i will approach j only if he prefers j to his current partner ($j \succ^i x^i$) and j is more prestigious than his current partner ($x^i \trianglerighteq j$). This concept is closely related to the abstract equilibrium concept discussed in Richter and Rubinstein (2015). Note that every C-equilibrium is an S-equilibrium.

A different interpretation of the S-equilibrium notion views a match as a kind of double ownership. When agents A and B are matched, A owns B, and simultaneously, B owns A! Agents are ranked according to some notion of value whereby each agent

"owns" his partner, and can "exchange" him for any weakly "less expensive" agent. In an S-equilibrium, no agent wishes to do so.

Double ownership might sound strange at first glance. As an example, consider a street of identical duplexes where each resident owns one unit. If all units are the same, then each duplex can be viewed as a partnership, with the only distinction between units being who your neighbor is. Selling a unit means selling the right (or duty) to be someone's neighbor. In this respect, two individuals living in the same duplex are involved in double ownership.

We now show that every S-equilibrium pairing is Pareto-optimal and that the S-equilibrium notion is quite different from pairwise stability. While both satisfy Pareto-optimality, even the existence of one concept does not imply the existence of the other:

Claim 3: (i) Every S-equilibrium pairing is Pareto-optimal.
(ii) There is a roommate economy with an S-equilibrium but no pairwise-stable pairing.
(iii) There is a roommate economy with a pairwise-stable pairing but no S-equilibrium.

Proof: (i) Recall that all preferences are assumed to be strict. Let $\langle \trianglerighteq, (x^i) \rangle$ be an S-equilibrium, and (y^i) a Pareto-superior pairing. Let i_1 be an agent with the highest $\trianglerighteq$-ranked partner in the pairing (x^i). Agent i_1 is the "wealthiest". Thus, his partner x^{i_1} is i_1's first-best and therefore $y^{i_1} = x^{i_1}$. Let i_2 be an agent with the highest $\trianglerighteq$-ranked partner among $N - \{i_1, x^{i_1}\}$. Again, it must be that $y^{i_2} = x^{i_2}$. Repeating this argument $n/2$ times leads to the conclusion that $(y^i) = (x^i)$.
(ii) Let $N = \{1,2,3,4\}$. The following is Gale and Shapley (1962)'s canonical example of a roommate problem without a pairwise-stable pairing:[5]

Agent	1	2	3	4
1st Preference	2	**3**	1	**1**
2nd Preference	3	1	**2**	2
3rd Preference	**4**	4	4	3

Table 1: A roommate problem with an S-equilibrium but no pairwise-stable pairing.

There are two S-equilibrium pairings: $\{1\leftrightarrow 4, 2\leftrightarrow 3\}$ (see Table 1) and $\{1\leftrightarrow 3, 2\leftrightarrow 4\}$, both supported by the ordering $1 \rhd 2 \rhd 3 \rhd 4$. The other pairing $\{1\leftrightarrow 2, 3\leftrightarrow 4\}$ is not an S-

[5]Consider any pairing. Let i be the agent matched with 4. Agent i prefers every other agent to 4 and there is $j \in \{1,2,3\}$ who top-ranks i. Thus, the pair (i, j) blocks the profile from being pairwise-stable.

equilibrium pairing under any ordering $\trianglerighteq$ since it must be that $3 \rhd 1$ (to prevent 2 from approaching 3), and it must be that $1 \rhd 3$ (to prevent 4 from approaching 1).

(iii) Consider the following example with $N = \{1,2,3,4\}$:

Agent	1	2	3	4
1st Preference	2	3	4	1
2nd Preference	**3**	**4**	**1**	**2**
3rd Preference	4	1	2	3

Table 2: A roommate problem with a pairwise stable pairing but no S-equilibrium.

The pairing $\{1\leftrightarrow 3, 2\leftrightarrow 4\}$ is the unique pairwise-stable pairing (see Table 2). We now show that no S-equilibrium exists: given any ranking $\trianglerighteq$, one of the agents, and without loss of generality let it be 1, is matched with his first-best and the resulting pairing is $\{1\leftrightarrow 2, 3\leftrightarrow 4\}$. However, it must then be that both $3 \rhd 1$ (to prevent 2 from approaching 3) and $1 \rhd 3$ (to prevent 4 from approaching 1). ■

Common Ranking: Recall that in the common-ranking roommate economy there is a unique C-equilibrium and wealth basically aligns with popularity. In contrast, any pairing is an S-equilibrium outcome, and it can be that an agent's "wealth" (in terms of the "market value" of his partner) is orthogonal to his own popularity. In particular, the pairing $\{i_1\leftrightarrow i_n, i_2\leftrightarrow i_{n-1}, \ldots\}$ together with the ranking $i_1 \rhd i_2 \rhd \cdots \rhd i_{n-1} \rhd i_n$ is an S-equilibrium in which the least-popular agent, i_n, is the "wealthiest" as he has the most prestigious partner.

Reciprocated Love: Here, the S-equilibrium pairing is unique, just like the C-equilibrium pairing, each agent ends up with his soulmate.

4. Jungle equilibria

Besides taboos and status, another force that governs societies is power (see Piccione and Rubinstein (2007)'s discussion of the Jungle Economy). By "power", we are not referring exclusively to raw physical strength, but also to gentler and subtler forms of power, such as seniority, conversational ability, or charm. In our model, power is represented by a strict ordering $\rhd$ over the agents where $a \rhd b$ means that a is more powerful than

b. Power limits the ability of agents to approach one another. The most obvious limitation is that a weaker agent may find it impossible to approach a stronger one, which is the basis of our J1-equilibrium notion. But power can be more intricate. In the J2-equilibrium, an agent is only able to approach another if both the approached agent and the approached agent's partner are weaker. The J3-equilibrium takes this a step further, stipulating that the approaching agent also has to be stronger than his current partner whom he seeks to abandon.

4.1 J1-equilibrium

A *J1-equilibrium* is a tuple $\langle \rhd, (x^i) \rangle$ for which there are no two agents, i and j, such that i is stronger than j and i strictly prefers j over his current partner x^i. That is, an agent is deterred from approaching another not necessarily due to a fear of rejection (as in the case of pairwise stability), but rather by the power of the desired agent.

The J1-equilibrium is a weaker concept than the C-equilibrium concept: if $\langle \trianglerighteq, (x^i) \rangle$ is a C-equilibrium, then $\langle \blacktriangleright, (x^i) \rangle$ is a J1-equilibrium, where $\blacktriangleright$ is any strict tie-breaking of $\trianglerighteq$. We now show that any J1-equilibrium outcome is pairwise stable (and consequently, Pareto optimal). However, the J1-equilibrium concept is more stringent, and may fail to exist even when a pairwise-stable pairing exists.

Claim 4: (i) Every J1-equilibrium pairing is pairwise stable.
(ii) A J1-equilibrium need not exist even when a pairwise-stable pairing exists.

Proof: (i) Let $\langle \rhd, (x^i) \rangle$ be a J1-equilibrium. Suppose that there are two agents i and j who strictly prefer each other to their current partners. Then the stronger agent prefers the weaker one over his current partner, violating the J1-equilibrium condition.
(ii) In the example from Table 2, the unique pairwise-stable pairing is $\{1 \leftrightarrow 3, 2 \leftrightarrow 4\}$ and by (i) this is the only J1-equilibrium candidate. However, it is not a J1-equilibrium pairing since the strongest agent has to get his first best and in $\{1 \leftrightarrow 3, 2 \leftrightarrow 4\}$ no agent does. ■

Common Ranking: In the common-ranking roommate economy, there is a unique J1-equilibrium pairing $\{i_1 \leftrightarrow i_2, i_3 \leftrightarrow i_4, \cdots, i_{n-1} \leftrightarrow i_n\}$ and the power ranking is any that satisfies $i_1, i_2 \rhd i_3, i_4 \rhd \cdots \rhd i_{n-1}, i_n$. Here's why: Since every agent desires i_1, he must be stronger than everyone else (except perhaps his partner), which means that i_1 has to be matched with his first-best, namely i_2. Since everyone else desires i_2, it must be that he

is stronger than them. Thus, i_1 and i_2 are the two most powerful agents. This pattern continues down the ranking. Among the remaining agents, i_3 and i_4 are matched and must be the next two most powerful agents, and so on. Thus, strength and popularity are essentially aligned in any J1-equilibrium. This contrasts with the S-equilibria for this economy, where any pairing is an S-equilibrium outcome, and wealth and popularity can be entirely at odds.

Reciprocated Love: Here, the J1-equilibrium pairing is unique, and every agent ends up with their soulmate. This is because between any two soulmates, one must be stronger than the other, enabling the stronger one to dictate the match. The power ranking $\rhd$ is completely arbitrary, as in the S-equilibrium case.

4.2 J2-equilibrium

A *J2-equilibrium* is a tuple $\langle \rhd, (x^i) \rangle$ in which there are no two agents i and j such that i desires j more than his current partner ($j \succ^i x^i$) and is more powerful than both j and j's partner ($i \rhd j$ and $i \rhd x^j$). Now an agent is deterred not only by the strength of the approached agent, but also by the strength of the approached agent's partner. Obviously, every J1-equilibrium is also a J2-equilibrium.

Claim 5: (i) A J2-equilibrium always exists.
(ii) Every J2-equilibrium pairing is Pareto-optimal.
(iii) There can be a Pareto-optimal pairing that is not a J2-equilibrium pairing.
(iv) Every S-equilibrium pairing is a J2-equilibrium pairing.

Proof: (i) Choose an arbitrary agent i_1 and make him the strongest agent. Pair him with his first-best, i_2, and make i_2 the weakest agent. Continue in this manner to obtain a sequence of agents such that $i_1 \rhd i_3 \rhd \cdots \rhd i_4 \rhd i_2$ and for each odd k, pair agent i_k with i_{k+1}. Notice that all odd-indexed agents are stronger than all even-indexed ones.

This algorithm achieves a J2-equilibrium: For every even k, agent i_k cannot approach any other agent because every couple has a member stronger than himself. For every odd k, if i_k prefers j over his partner, then j must have been removed in the above construction before i_k, and therefore, either j or j's partner is stronger than i_k.

(ii) Let $\langle \rhd, (x^i) \rangle$ be a J2-equilibrium. Assume that (y^i) Pareto-dominates (x^i). Let j be the strongest agent in $M = \{i : x^i \neq y^i\}$. Then $y^j \succ^j x^j$, and j is stronger than both y^j and y^j's original partner, x^{y^j} (because both are in M), which violates the J2-equilibrium definition.

(iii) Consider again the profile of agents' preferences from Table 2. The pairing $\{1 \leftrightarrow 3, 2 \leftrightarrow 4\}$ is Pareto-optimal. If it were a J2-equilibrium outcome, then the strongest agent would be matched to his first best, but no agent is matched with his first-best partner.

(iv) Let $\langle \trianglerighteq, (x^i) \rangle$ be an S-equilibrium and break ties so that $\trianglerighteq$ is strict. Define a power ranking $\blacktriangleright$ by ranking agents according to the status of their partners: $i \blacktriangleright j$ if $x^i \rhd x^j$. Suppose that for some i and j, it holds that $j \succ^i x^i$. Then, it must be that $j \rhd x^i$ since $\langle \trianglerighteq, (x^i) \rangle$ is an S-equilibrium. But then $x^j \blacktriangleright i$. Therefore, $\langle \blacktriangleright, (x^i) \rangle$ is a J2-equilibrium. ■

Common Ranking: Recall that in this economy every pairing is an S-equilibrium pairing. By Claim 5(iv) it follows that all pairings are J2-equilibrium pairings. This is in stark contrast to the uniqueness of the J1-equilibrium pairing.

While every pairing is part of some J2-equilibrium, this is not the case for every power relation. For example, when $n = 4$, there is no J2-equilibrium with the power relation $i_4 \rhd i_1 \rhd i_2 \rhd i_3$. This is because i_4 is the most powerful, and must be matched with i_1 who is everyone's first choice. But then, the candidate pairing is $\{i_1 \leftrightarrow i_4, i_2 \leftrightarrow i_3\}$ which is not a J2-equilibrium since i_1 prefers i_2, and is stronger than both i_2 and i_3.

Reciprocated Love: Here again, the J2-equilibrium pairing is unique, each agent ends up with his soulmate. The strongest agent must be matched with his soulmate, and then the strongest among the remaining agents is matched with his soulmate (who is still remaining), and so on. Again, the power ranking $\rhd$ is arbitrary.

4.3 J3-equilibrium

A *J3-equilibrium* is a tuple $\langle \rhd, (x^i) \rangle$ in which there are no two agents i and j such that i desires j more than his current partner ($j \succ^i x^i$) and is more powerful than j, j's partner and his own partner ($i \rhd j, x^i, x^j$). That is, for agent i to approach another agent j, he now also needs to be strong enough to leave his current partner.

Clearly, every J2-equilibrium is a J3-equilibrium. By Claim 6 below, the set of J3-equilibrium pairings is identical to those of the J2-equilibrium, but unlike the case of J2-equilibrium, every power relation is a part of some J3-equilibrium.

Claim 6: (i) For every power relation $\rhd$, there is a unique J3-equilibrium $\langle \rhd, (x^i) \rangle$.
(ii) The set of J3-equilibrium pairings is identical to the set of J2-equilibrium pairings.

Proof: (i) Fix a power relation $\rhd$. We now construct a J3-equilibrium with this power relation using an adapted serial dictatorship algorithm. In the first step, the $\rhd$-strongest agent picks his most-preferred partner. Both agents are removed. In each subsequent step, the $\rhd$-strongest remaining agent chooses his most-preferred partner from among those remaining and both are removed. By this algorithm, half of the agents "make a choice" while the other half "are chosen". To see that this is a J3-equilibrium, notice that any agent who "makes a choice" can only prefer earlier removed agents, i.e. those who are stronger than him or are paired with a stronger partner than him. Any agent who does not make a choice is neutralized from approaching any other agent because he is matched with a stronger agent.

For uniqueness, suppose there are two different J3-equilibria with the same power relation, $\langle \rhd, (x^i) \rangle$ and $\langle \rhd, (y^i) \rangle$. Let j be the $\rhd$-strongest agent in $M = \{i : x^i \neq y^i\}$. WLOG, suppose that $x^j \succ^j y^j$. The tuple $\langle \rhd, (y^i) \rangle$ is not a J3-equilibrium because j prefers x^j to y^j, and $j \rhd y^j, x^j, y^{x^j}$ since y^j, x^j and y^{x^j} are all in M.

(ii) As mentioned, every J2-equilibrium is also a J3-equilibrium. It remains to be shown that any J3-equilibrium pairing is a J2-equilibrium pairing (possibly with a different power relation). Consider a J3-equilibrium $\langle \rhd, (x^i) \rangle$. In every couple, there is a stronger agent and a weaker one. Let S be the set of stronger agents and W be the set of weaker ones. Define a new power relation $\rhd'$ by preserving $\rhd$ on S and on W and pushing all members of W below all members of S. The tuple $\langle \rhd', (x^i) \rangle$ is a J2-equilibrium: Suppose not, namely, there are i and j such that $j \succ^i x^i$ and $i \rhd' x^j, j$. Suppose $j \rhd x^j$ and thus $j \in S$ (the case of $x^j \rhd j$ is analogous). Since $i \rhd' j$, it must also be that $i \in S$. Thus, (1) $i \rhd x^i$ (as $i \in S$), (2) $i \rhd j$ (since $\rhd = \rhd'$ on S and $i, j \in S$) and (3) $i \rhd x^j$ (as $i \rhd j$ and $j \rhd x^j$ since $j \in S$). This contradicts $\langle \rhd, (x^i) \rangle$ being a J3-equilibrium. ■

To construct a J3-equilibrium, the proof of Claim 6 uses an adapted serial dictatorship algorithm. The algorithm is adapted in the sense that rather than choosing over objects or bundles, the objects of choice are the agents themselves, and if an agent is chosen, then he loses his ability to choose (whether or not he would indeed choose his current partner). We do not further analyze our running examples here because the J2- and J3-equilibrium pairings are identical.

4.4 External vs. internal power relations

In the original jungle equilibrium concept of Piccione and Rubinstein (2007), the power relation between agents is external. In contrast, in our definitions of jungle equilibria (and as in Rubinstein and Yıldız (2022)), power is determined as part of the equilibrium concept, which is akin to the endogeneity of prices in the standard Walrasian setting.

Note that not all power relations are consistent with the J1- or J2-equilibrium concepts, just as not all price vectors constitute a Walrasian equilibrium. On the other hand, regarding the J3-equilibrium concept, any power relation corresponds to a unique equilibrium outcome, and therefore this solution concept fits both internal and external power structures.

5. Final comments

5.1 Two types of equilibria

The equilibrium concepts we discuss in this paper fall into two categories: those resembling market equilibrium, where agents face "budget sets" and the optimal *choice* profile must be feasible, and those that are more akin to non-cooperative solutions, where an equilibrium is a feasible profile that is immune to certain classes of *unilateral deviations*. Let us elaborate.

In a standard competitive equilibrium, every agent possesses an endowment, a price system is endogenously determined, agents make optimal choices based on their endowment and the prices without regard to the choices of others, and their choices are coherent. Our T-, C-, and J1-equilibrium concepts align with this framework. In these concepts, an additional price-like element – either the set of permissible alternatives,

status ranking, or power ranking – is formed endogenously. This element dictates what options agents can choose from (their "budget sets") and the agents select their options according to their preferences, without regard to the choices of others. In equilibrium, the profile of choices is feasible.

On the other hand, the S-, J2-, and J3-equilibrium concepts are predicated on immunity to deviations (as in non-cooperative game theory). In these concepts, a candidate pairing constitutes an equilibrium if it is immune to a class of unilateral deviations which is determined by an endogenous price-like element (either a status or power ranking). For these concepts (and pairwise stability as well), the set of deviations that agents can make depends upon the equilibrium configuration, unlike in the T- and C-equilibrium settings.

5.2 Relationship between the solution concepts' outcomes

The central distinction of the current paper from the prior literature on pairwise stability is that deviations are now unilateral (as in Nash equilibrium) rather than coalitional (as in pairwise stability).

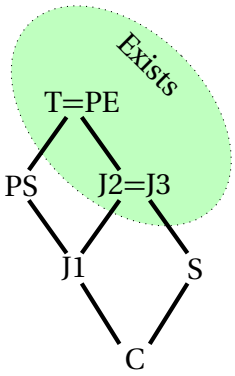

Figure 1: Inclusion relationship between the equilibrium notions, pairwise stability (PS) and Pareto efficiency (PE). J2-, J3- and T-equilibria always exist.

Figure 1 illustrates the connections between the various equilibrium concepts introduced in this paper, namely T, S, C, J1, J2, and J3, and how they relate to Pareto-optimality and pairwise stability.[6] Existence always holds for the J2-, J3- and T-equilibrium concepts. For all of them, the First Welfare Theorem holds. The Second Welfare Theorem is guaranteed only for the T-equilibrium.

5.3 Algorithms

One of the central attractions of pairwise stability for the two-population matching problem is that Gale and Shapley (1962) demonstrate that the "deferred acceptance" algorithm achieves either the male- or female-optimal pairwise-stable pairing. However, the algorithm only works for two-population models, because the two sides are treated asymmetrically, and therefore cannot be applied to the roommate problem. Some of the literature on the roommate problem proposes algorithms that yield a pairwise-stable profile when one exists (although often it does not). We have proposed constructive algorithms for the C-, J2-, and J3- equilibrium concepts. We find the adapted serial dictatorship algorithm, used in Claim 6 to construct a J3-equilibrium pairing, to have interesting features: (i) some agents are dictators while others are dictated to and (ii) any J3-equilibrium pairing can be obtained by adjusting the parameters of the algorithm (a power ordering).

5.4 Unilateral stability vs. pairwise stability

As mentioned in the introduction, we do not argue that any of the proposed solution concepts is superior or inferior to pairwise stability, on either normative or positive grounds. We have intentionally avoided any speculation as to which of the examined institutions might be more likely to emerge. Rather, the proposed solution concepts express fundamentally different concerns for harmony in a society, with each concept narrating its own story.

[6] All of the depicted relationships between the equilibrium cases have been demonstrated in the previous claims except for the non-inclusion of the S- and J1-equilibrium notions. Table 2 presents a case with an S-equilibrium but no J1-equilibrium. It is easy to formulate an example which shows the reverse.

References

Alcalde, José (1995). "Exchange-Proofness or Divorce-Proofness? Stability in One-Sided Matching Markets". *Economic Design, 1*, 275–287.

Gale, David and Lloyd Shapley (1962). "College Admissions and the Stability of Marriage". *The American Mathematical Monthly, 69*, 9–15.

Morrill, Thayer (2010). "The Roommates Problem Revisited". *Journal of Economic Theory, 145*, 1739–1756.

Piccione, Michele and Ariel Rubinstein (2007). "Equilibrium in the Jungle". *The Economic Journal, 117*, 883–896.

Richter, Michael and Ariel Rubinstein (2015). "Back to Fundamentals: Equilibrium in Abstract Economies". *American Economic Review, 105*, 2570–2594.

Richter, Michael and Ariel Rubinstein (2020). "The Permissible and the Forbidden". *Journal of Economic Theory, 188*, article 105042.

Root, Joseph and David S. Ahn (2020). "Incentives and Efficiency in Constrained Allocation Mechanisms". mimeo.

Rubinstein, Ariel and Kemal Yıldız (2022). "Equilibrium in a Civilized Jungle". *Theoretical Economics, 17*, 943–953.

Tan, Jimmy J. M. (1990). "A Maximum Stable Matching for the Roommates Problem". *BIT Numerical Mathematics, 30*, 631–640.