Dynamic Games and Applications (2023) 13:462–479
https://doi.org/10.1007/s13235-022-00474-8

Farsighted Rationality in Hedonic Games

G.-Herman Demeze-Jouatsa[1] · **Dominik Karos**[1]

Accepted: 7 August 2022 / Published online: 5 September 2022

Abstract
We consider a hedonic coalition formation game in which a coalition chooses for each partition of the player set the probability with which it forms and thereby destroys the current partition. These probabilities are commonly known so that farsighted players know at every partition what future partitions, and hence payoffs, will be reached with what probability. Thus, players can make rational decisions about the moves they support. We show that if coalitions make mistakes with small but positive probability, then there is a behavior profile in which no coalition has a profitable one-shot deviation.

Keywords Abstract games · Hedonic games · Farsighted stability · Coalition stable equilibrium

JEL Classification C71 · C72

1 Introduction

A hedonic game [8] specifies for each player a payoff in each coalition, giving players preferences over coalitions. In particular, unlike in games in characteristic function form, there is no competition over payoffs within any coalition: If a coalition forms, each player's payoff is determined. When solving a hedonic game, one is, hence, only interested in how players partition into coalitions. Arguably, the most prominent solution is the set of *core*

This article is part of the topical collection "Group Formation and Farsightedness" edited by Francis Bloch, Ana Mauleon and Vincent Vannetelbosch.

Comments from the IMW discussion group as well as the anonymous referees are gratefully acknowledged. G.-Herman Demeze-Jouatsa: The author gratefully acknowledges financial support from the DFG (Deutsche Forschungsgemeinschaft/German Research Foundation) via grant Ri 1128-9-1 (Open Research Area in the Social Sciences, Ambiguity in Dynamic Environments) .

✉ Dominik Karos
dominik.karos@uni-bielefeld.de

G.-Herman Demeze-Jouatsa
demeze_jouatsa@uni-bielefeld.de

[1] Center for Mathematical Economics, Bielefeld University, Bielefeld, Germany

stable partitions: In such a partition no set of players can increase their payoffs by forming a new coalition. Unfortunately, despite their rather simple structure, hedonic games might not have core stable partitions. [4] and [2] provide sufficient conditions for the nonemptiness of the core; [12] provides a condition which is both necessary and sufficient and very similar to the balancedness condition by [16] and [5].

The idea of the core, namely that there be no profitable formation of new coalitions, is quite intriguing; yet, it requires the assumption of players' naivety: When a coalition considers forming, all its members compare their current payoff with their payoff after forming. No attention is being paid to future deviations by some of its members, and no attention is being paid to other coalitions who might leave the status quo. That is, players are *myopic*.

The first investigation of *farsighted* players in hedonic games has been provided by [7] who used the more general models of [9] and [6]. Although their analysis captures players' rationality in that they can anticipate the consequences of their deviating, there are two potentially problematic assumptions: First, it is assumed that whenever a player deviates from a coalition, the remainder of this coalition stays intact. Second, players compare payoffs from all reasonable future deviations to the status quo. So, they might act together even though their final goals differ (see, for instance, [3]), and they do not take into account that other coalitions might move preemptively. The result is a solution which always exists, but which is too permissive and which does not account for the *full* rationality of players.

In this paper, we introduce a different way to talk about farsighted players in hedonic games, which is based on [15] and [13]. For that purpose, we translate a hedonic game into an abstract game. This abstract game considers the set of partitions as state space and specifies for any partition and coalition what new partition emerged if this coalition formed. We provide four axioms for this specification that ensure that all players have the same unique expectation about potential moves among partitions.

The abstract game describes a coalition formation game that is similar to [14], where coalitions are endowed with strategies (behaviors) that specify at each partition whether or not to form (if they are not already part of that partition). A behavior profile, thus, defines transitions among partitions, which in turn define a Markov process. The stationary distribution of such a process determines how much time is spent in each partition. Thus, the payoff from any behavior profile is a weighted average of payoffs, where the weights are given by the relative time spent in each partition.

In a *weak equilibrium*, each coalition behaves optimally in each partition π, given the behavior of all coalitions (including itself) at all other partitions, and the behavior of all other coalitions at π. Thus, a weak equilibrium is stable with respect to *one-shot deviations*. In contrast to [14] we allow coalitions to play mixed strategies; in fact, we restrict our analysis to strategies that play each pure behavior with some small but positive probability $\varepsilon > 0$, taking into account the possibility to make mistakes. This ensures that there is a path of moves between any two partitions, so that any partition is reached with positive probability. Thus, any optimality condition in equilibrium applies to all partitions in which a coalition might have to decide whether or not to form.

The main result of our paper is that for every $\varepsilon > 0$ a weak equilibrium exists. The mathematical difficulty in showing this result is that players' payoff functions are not linear in the probability with which each behavior is being played. Thus, showing that the set of best replies is convex (as it is in normal form games) is difficult. (If we allow for deviations that are not one-shot, then the set of best replies is, in fact, not convex.) But once, convexity is proven, obtaining the result is straightforward.

The remainder of the paper is structured as follows: In Sect. 2, we introduce the necessary notation, recall the definition of hedonic games and introduce hedonic coalition formation

games that are a special class of abstract games. In Sect. 3 we introduce coalition behaviors, translate them into transition matrices and derive the relevant payoff functions. Section 4 introduces the equilibrium and proves its existence. We close the paper with Sect. 5 where we show that weak equilibria are not necessarily stable with respect to arbitrary deviations.

2 Preliminaries

2.1 Hedonic Games

Let N be a finite set of players. Subsets $S \subseteq N$ are called *coalitions*. For $S \subseteq N$ write 2^S for the set of subsets of S, and $P(S)$ for the set of nonempty subsets. A *partition* is a collection $\pi = \{S^1, \ldots, S^m\}$ of nonempty coalitions such that $\bigcup_{k=1}^m S^k = N$ and $S^k \cap S^l = \emptyset$ for all $k \neq l$; the set of all partitions is denoted by Π. For $i \in N$ and a partition π we write $\pi(i)$ for the unique element of π that contains i. A *hedonic game* is a map v that maps each nonempty coalition S to some $v(S) \in \mathbb{R}^S$. That is, a hedonic game is a cooperative game such that each player's payoff in each coalition is uniquely determined: There is no negotiation over payoffs within coalitions whatsoever. For any hedonic game v, we define the map $V : \Pi \to \mathbb{R}^N$ by $V_i(\pi) = v_i(\pi(i))$. That is, $V(\pi) \in \mathbb{R}^N$ is the payoff vector if partition π forms.

Arguably, the most prominent solution of a hedonic game is the set of *core partitions*: Those partitions for which no coalition has an incentive to deviate. To make this precise, we say that a partition π is *dominated* via S if $v_i(S) > V_i(\pi)$ for all $i \in S$. The *core* is the set of undominated partitions.[1]

Example 2.1 (**The roommate problem**) There are three players who have to decide about who of them will be moving in together in a two-bedroom flat. They have somewhat conflicting interests: While everybody dislikes to move in with three people into a two-bedroom flat, 1 prefers to move in with 2 over moving in with 3 over staying alone; 2 prefers moving in with 3 over moving in with 1 over staying alone; and 3 prefers moving in with 1 over moving in with 2 over staying alone. Suppose that payoffs from staying alone are 0, from moving into an overcrowded place is -1, from getting the preferred roommate is 4, and from getting the other room mate is $a \in (0, 4)$. This game does not have a core stable outcome: Surely, neither the partition into singletons nor the partition that only contains the grand coalition are core stable. But neither are the others: Whenever two players have formed a coalition, one of the two has an incentive to form a new one with the outside player.[2] □

The core is a myopic concept: At any partition π, the members of a potential coalition S compare their payoffs from forming with those at π. No attention is paid to any moves other coalitions (or even some members of S) could make after S has formed. In particular, the players in S do not take the behavior of those in $N \setminus S$ into account: It is irrelevant for $v(S)$, and S operates under the presumption that no one will react upon their deviation.

If players are not myopic, they will account for the possibility that after their own deviation other coalitions might form. Thus, they have to make assumptions about what happens to those "left behind." So, a dominance relation cannot simply be defined between a partition and a coalition, but rather between two partitions. [7] define such a dominance relation based on

[1] As the payoffs of all players are determined by a partition π, there is no need to explicitly consider the set of core payoff vectors.

[2] Recently, [1] have proposed a solution to the roommate problem that is based on the credibility of deviations. They show, in particular, that if one allows for "weak" deviations, then existence is guaranteed.

[6]: Partition π' farsightedly dominates partition π if there is a sequence of pairs $\left(S^l, \pi^l\right)_{l=1}^m$ such that

$$\pi^l = \left\{S^l\right\} \cup \left\{T \setminus S^l\right\}_{T \in \pi^{l-1}} \setminus \{\emptyset\}$$

for $l = 1, \ldots, m$, where $\pi^0 = \pi$, $\pi^m = \pi'$, and $V_i\left(\pi'\right) > V_i\left(\pi^{l-1}\right)$ for all $i \in S^l$. A solution that is based on such a definition seems, at least from a rationality point of view, more plausible than a purely myopic solution. Still, there are some caveats: For instance, it makes the implicit assumption that players who are left behind stay together. Another problem is that different members of a coalition S might only work together because they have different, and potentially contradicting, expectations about how the game unfolds.[3] The most severe issue, however, seems to be that coalitions still do not behave rationally: They make assumptions about the consequences of their forming, but they ignore the consequences of their not forming.

In order to overcome these issues, we shall translate hedonic games into abstract games for which farsightedness has recently gained some attention. In general, an *abstract game* is a tupel $\left(N, X, (\rightarrow_S)_{S \in P(N)}, (U_i(\cdot))_{i \in N}\right)$, where X is a set of states, $U_i : X \rightarrow \mathbb{R}$ is player i's utility function over states and $\rightarrow_S$ describes coalition S's ability to move from one state to another: For two states $x, y \in X$ we write $x \rightarrow_S y$ if S can replace x with y. In this case, we say S is *effective* for a move from x to y. In the context of a hedonic game we choose $X = \Pi$, i.e., the set of states is exactly the set of partitions, and $U_i = V_i$, which is i's payoff function over partitions.

Example 2.2 Recall the roommate problem in Example 2.1. A potential $\rightarrow$ for this hedonic game in depicted in Fig. 1, where $\pi^0 = \{\{1\}, \{2\}, \{3\}\}, \pi^1 = \{\{1, 2\}, \{3\}\}, \pi^2 = \{\{1\}, \{2, 3\}\}$, $\pi^3 = \{\{2\}, \{1, 3\}\}$, and $\pi^4 = \{N\}$. □

As the profile $\rightarrow = (\rightarrow_S)_{S \in P(N)}$ is the most relevant piece of the puzzle, we shall have a closer into it look in the next section.

2.2 Effectivity in Hedonic Coalition Formation

The question of what partitions can arise in a hedonic game hinges on the coalitions' abilities to change partitions. The hedonic game itself remains quite agnostic about this as it only specifies payoffs for coalitions and nothing more. Thus, we shall derive four assumptions on coalitions' abilities to affect partitions, which are reflected in $\rightarrow$. First, we would expect $\rightarrow$ to satisfy:

H1 If $\pi \rightarrow_S \pi'$, then $S \in \pi'$.

That is, S can only move from a partition π to a partition π' if it is a member of the latter. Observe that we do not allow the members of S to jointly form a partition of S: If they collaborate, they must form a coalition. As [15] point out, the action of farsighted players in S depends on the expected reaction by $N \setminus S$ as this might influence future deviations. To avoid unintuitive results, they propose two conditions and refer to them as *coalition sovereignty*[4]:

H2 If $\pi \rightarrow_S \pi'$, $T \in \pi$, and $S \cap T = \emptyset$, then $T \in \pi'$.

[3] A similar issue arises in [11], who consider "robust" deviations in the roommate problem. A deviation is robust up to depth k if none of the deviators will be worse off after any sequence of at most k subsequent deviations than at the original partition.

[4] [15] formulate their conditions for general NTU games; we provide here the adaption to hedonic games.

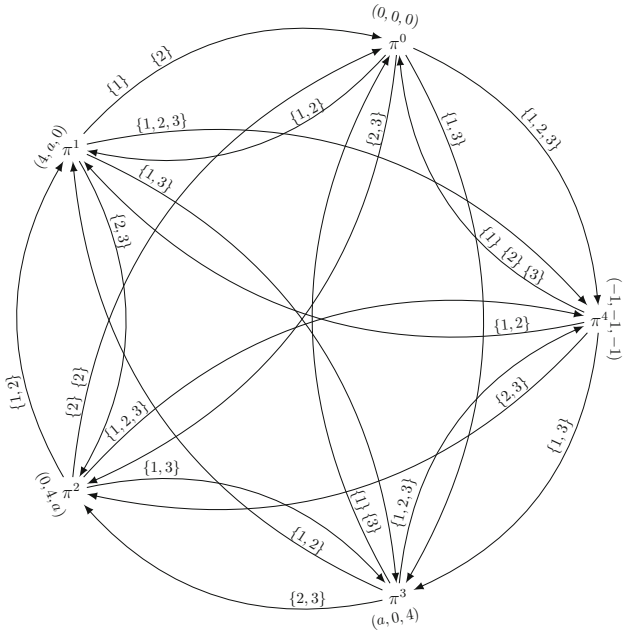

Fig. 1 The roommate problem

H3 For every $\pi \in \Pi$ and $S \in P(N)$ there is π' with $S \in \pi'$ such that $\pi \to_S \pi'$.

Condition **H2** requires that a coalition S that deviates from partition π has no influence over coalitions that have not been affected by its deviation. That is, a coalition in π that did not intersect with S will not change. Condition **H3** requires that from each partition π each coalition S that is not a member of π can deviate. Both conditions are highly appropriate in the context of hedonic games: They endow coalitions with the power to form at any state, yet they ensure that no coalition has the power to affect the behavior of others when moving.[5] An observation worth making is that **H1** and **H2** together imply $\pi = \pi'$ whenever $S \in \pi$ and $\pi \to_S \pi'$.

The transition between partitions in [7] satisfies Conditions **H1** – **H3**, but these conditions alone still allow for quite a range of partitions π' that a coalition S might move to from π, as nothing has been said about those players who were "left behind" by S. Define for any partition π and any coalition S the set $\pi(S)$ by $\pi(S) = \bigcup_{i \in S} \pi(i)$, which is the set of all players whose coalitions are affected by a deviation of S. There is no reason to presume S have power about the behavior of $\pi(S) \setminus S$. Yet, we shall assume that there is a (common) expectation about their behavior. A *residual map* is a map τ, which maps each pair (π, S) on a partition $\tau\ (\pi, S)$ of the set $\pi(S)$ with $S \in \tau\ (\pi, S)$. For $i \in \pi(S)$ we write $\tau\ (i \mid \pi, S)$ for the unique element of $\tau\ (\pi, S)$ that contains i.

H4 There is a residual map τ such that if $\pi \to_S \pi'$, then $\pi'(i) = \tau(i \mid \pi, S)$ for all $i \in \pi(S)$.

[5] This is not to say that there are no later moves that such groups might want to undertake, or that such moves are not being expected. Such moves, however, are the decisions of the moving groups at the new partition rather than a decision of the deviating coalition at the old one.

Condition **H4** ensures that the behavior of $\pi(S)$ cannot be chosen by S, yet is uniquely determined and commonly known. Thus, **H2** and **H4** together simply ensure that all coalitions have a common expectation about the immediate consequences of any move.

We shall not impose any conditions on the residual map; for the remainder its existence is sufficient. Yet, there are several instance of τ that have been investigated in the literature before. For instance, [10] consider two variants: The γ-model, where coalitions who are left behind split up into singletons, and the δ-model, where they remain as they were.

Example 2.3 For any pair (π, S) of a partition and a coalition let $\gamma(\pi, S) = \left\{S, \{\{i\}\}_{i\in\pi(S)\setminus S}\right\}$. The unique $\to^{\gamma}$ that satisfies **H1**–**H4** with $\tau = \gamma$ is

$$\pi \to_S^{\gamma} \pi' \quad \text{if and only if} \quad \pi' = \{S\} \cup \{T\}_{T\in\pi\setminus\pi(S)} \cup \{\{i\}\}_{i\in\pi(S)\setminus S}.$$

For any pair (π, S) of a partition and a coalition let $\delta(\pi, S) = \left\{S, \{\pi(i) \setminus S\}_{i\in\pi(S)\setminus S}\right\} \setminus \{\emptyset\}$. The unique $\to^{\delta}$ that satisfies **H1**–**H4** with $\tau = \delta$ is[6]

$$\pi \to_S^{\delta} \pi' \quad \text{if and only if} \quad \pi' = \{S\} \cup \{T \setminus S\}_{T\in\pi} \setminus \{\emptyset\}.$$

Observe that $\to$ in Example 2.2 corresponds to the γ-model. This can be seen from $\pi^4 \to_{\{i\}} \pi^0$ for $i = 1, 2, 3$. □

We allow coalitions to act strategically when deciding whether or not to move, and we are interested in their equilibrium behavior. We assume that in each partition π coalitions are allowed to move in a specified order that is described by a bijection $\rho^{\pi} : \left\{1, \ldots, 2^{|N|} - 1\right\} \to P(N)$. Here, $\rho^{\pi}(l)$ is the l-th coalition that is allowed to move at π.[7]

Definition 2.4 A *hedonic coalition formation game* is a tuple $(N, V, \to, \rho)$, where V is the payoff function from a hedonic game with player set N, $\to$ satisfies conditions **H1**–**H4**, and $\rho = (\rho^{\pi})_{\pi\in\Pi}$ is an order profile.

As **H2** and **H4** together uniquely determine the behavior of $N \setminus S$ for any S and π, we obtain the following result. Its proof, as all proofs, can be found in the appendix.

Theorem 2.5 *Let $(N, V, \to, \rho)$ be a hedonic coalition formation game. Then, for each partition π and each coalition $S \subseteq N$ there is a unique partition π' with $\pi \to_S \pi'$.*

Observe that if a coalition S could decide to form a partition of S (which would violate **H1**), then S could move to more than one other partition, and Theorem 2.5 would not hold.

3 Analysing Hedonic Coalition Formation Games

3.1 Coalition Behavior and Transitions

In a hedonic coalition formation game, any coalition (that has not formed yet) has only two options: To be or not to be? That is the question. The only strategic decision that a coalition has to make (at any partition) is, hence, to choose the probability with which to form.[8] Thus,

[6] This is the assumption that [7] use as well.

[7] There is nothing specific about having a deterministic order at each π, a random order would suffice, as long as the distribution over orders if commonly known.

[8] We presume history independence here: A coalition's decision at π only depends on π and not on how or when π was reached.

a *(mixed) coalition behavior*[9] of coalition S is a map $\beta_S : \Pi \to [0,1]$, where $\beta_S(\pi)$ denotes the probability that S forms and deviates from π.[10] We write $\Delta_S \subseteq [0,1]^\Pi$ for the set of all coalition behaviors of S. A *behavior profile* is a vector $\beta = (\beta_S)_{S\in P(N)} \in \Delta = \times_{S\in P(N)}\Delta_S$ of behaviors.

Example 3.1 Recall $\to$ for the 3-player roommate problem in Example 2.2 and consider the hedonic coalition formation game $(N, V, \to, \rho)$ where $\rho^\pi = \rho$ for all $\pi \in \Pi$ and ρ is defined by

$$\rho(1) = \{1\} \qquad \rho(2) = \{2\} \qquad \rho(3) = \{3\} \qquad \rho(4) = \{1,2\}$$
$$\rho(5) = \{2,3\} \qquad \rho(6) = \{1,3\} \qquad \rho(7) = \{1,2,3\}.$$

Consider the following behavior profile. $\beta_{\rho(l)}(\pi) = 0$ for all $\pi \in \Pi$ and $l = 1,2,3,7$. Further, $\beta_{\rho(l)}\left(\pi^0\right) = \beta_{\rho(l)}\left(\pi^4\right) = 1$ for $l = 4,5,6$. Lastly,

$$\beta_{\rho(4)}\left(\pi^2\right) = \beta_{\rho(5)}\left(\pi^3\right) = \beta_{\rho(6)}\left(\pi^1\right) = p,$$
$$\beta_{\rho(4)}\left(\pi^3\right) = \beta_{\rho(5)}\left(\pi^1\right) = \beta_{\rho(6)}\left(\pi^2\right) = \beta_{\rho(4)}\left(\pi^1\right) = \beta_{\rho(5)}\left(\pi^2\right) = \beta_{\rho(6)}\left(\pi^3\right) = 0,$$

where $p \in (0,1)$. That is, each pair would deviate from the grand coalition and from the singleton partition with probability 1; and whenever a pair has formed, another pair will deviate with probability p. Surely, both π^0 and π^4 will be left for π^1 by $\{1,2\}$ with probability 1. The probability that partition π^1 will be left is p, and if it is left, then by a move of $\{1,3\}$ to π^3. Similarly, π^2 will be left with probability p to π^1, and π^3 will be left with probability p to π^2. So, the transition probabilities between partitions are described by the $(\Pi \times \Pi)$-dimensional matrix

$$P^\beta = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 1-p & p & 0 & 1 \\ 0 & 0 & 1-p & p & 0 \\ 0 & p & 0 & 1-p & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where $P_{\pi',\pi}$ denotes the probability of a transition from π to π'. □

3.2 Markov Processes and Expected Payoffs

Given an order profile $\rho = (\rho^\pi)_{\pi\in\Pi}$ and a behavior profile β, the probability of a transition from π to π' is

$$P^\beta_{\pi',\pi} = \sum_{l:\pi\to_{\rho^\pi(l)}\pi'} \prod_{h<l} \left(1 - \beta_{\rho^\pi(h)}(\pi)\right) \beta_{\rho(l)}(\pi) \qquad \text{for all } \pi,\pi' \in \Pi, \pi' \neq \pi, \tag{1}$$

and the probability, that no coalition will move out of π is, hence,

$$P^\beta_{\pi,\pi} = \prod_{l=1}^{2^{|N|}-1} \left(1 - \beta_{\rho^\pi(l)}(\pi)\right) \qquad \text{for all } \pi \in \Pi. \tag{2}$$

[9] [14] defined a coalition behavior for a general abstract game $(N, X, \to, U)$ as a map $\beta_S : X \to X$ with $x \to_S \beta_S(x)$. Given the special form of hedonic coalition formation game and the observation in Theorem 2.5, our definition is equivalent to a mixed behavior in this sense.

[10] For convenience, if $S \in \pi$, then $\beta_S(\pi) = 0$.

If the behavior profile β is such that the Markov process with transition matrix P^β converges independently of its starting point towards a unique partition π, then the expected payoff from β is easily determined, namely $V(\pi)$. But in general, we cannot expect such a partition to exist; in fact, one of the main problem in hedonic games are circles among partitions such as in the roommate problem. In this case, we can define payoffs by asking: How much time will be spent in each partition given some behavior profile β? For this purpose note that if a partition π is reached, then after n periods of possible moves among states, the expected number of periods spent in some π' is given by the π'-th entry of the π-the column of the matrix $\sum_{m=1}^{n}\left(P^\beta\right)^m$. Thus, if the process starts at π and we do not impose any restrictions on the number of periods during which coalitions are allowed to move, then the expected relative amounts of time spent in each partition are given by the π-th column of the matrix $\lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n}\left(P^\beta\right)^m$, which is given by Π-dimensional vector

$$\mu^\beta = \lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n}\left(P^\beta\right)^m e^\pi, \tag{3}$$

where e^π is the Π-dimensional unit vector with 1 as its π-th entry. With a slight abuse of notation we shall write $\mu^\beta\left(\pi'\right)$ for the vector entry $\mu^\beta_{\pi'}$.[11] We shall formulate a condition on P^β such that μ^β does not depend on the choice of π. This is particularly important for the case of hedonic games as these games do not specify any initial partition.

The Markov process with transition matrix P^β is called *irreducible* if for every two partitions π, π' there is $m \in \mathbb{N}$ such that $\left(\left(P^\beta\right)^m\right)_{\pi',\pi} > 0$. The following proposition comprises well-known results about irreducible Markov processes with finite state space that we will need later. We do not provide a proof but refer the reader to the standard literature, e.g., [17].

Proposition 3.2 *Let P be the transition matrix of an irreducible Markov process over Π. Then there is a unique vector $\mu \in \mathbb{R}^\Pi$ with $\sum_{\pi\in\Pi}\mu(\pi) = 1$ that satisfies (3) for all $\pi \in \Pi$. In particular, $\mu\left(\pi'\right) > 0$ for all $\pi' \in \Pi$ and μ satisfies $P\mu = \mu$, i.e., μ is the unique eigenvector of P to eigenvalue 1 with length 1.*

Example 3.3 Recall the transition matrix in Example 3.1. The Markov process that is defined by the transition matrix P^β has stationary distribution $\left(0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0\right)$. This means that after a very long time it will have spent the same amounts of time in π^1, π^2, and π^3, while it will not have spent any time in π^0 or π^4. Observe that even if the game starts in π^0 or π^4, these partitions will be left in the first period any never be returned to. Thus, relative time spent there converges towards 0. □

We can now define payoffs for those behavior profiles β for which P^β is the transition matrix of an irreducible Markov process.

Definition 3.4 Let β be a behavior profile such that P^β be the transition matrix of an irreducible Markov process with the unique stationary distribution μ^β. Then the payoffs from the behavior profile β are

$$u_i(\beta) = \sum_{\pi\in\Pi}\mu^\beta(\pi)V_i(\pi) \tag{4}$$

for all $i \in N$. □

[11] Note that $\mu^\beta(\pi) \geq 0$ for all $\pi \in \Pi$ and $\sum_{\pi\in\Pi}\mu^\beta(\pi) = 1$, so that μ^β is indeed a probability distribution over Π.

Example 3.5 Recall the game in Example 2.2, the behavior in in Example 3.1 and the corresponding stationary distribution in Example 3.3. Here, the payoffs are given by $u_i(\beta) = \frac{4+a}{3}$ for all $i \in N$. □

While the payoff function in (4) is rather intuitive, it has two severe caveats: First, it is not well defined if the stationary distribution of P^β is not unique. Second, unlike payoff functions in standard normal form games, it is not linear in β. Thus, when looking for an equilibrium that is based on (some form of) best replies, it is not trivial to show that the set of best replies is convex.

3.3 Errors and ε-Behaviors

We are interested in the following class of behavior profiles which lead to irreducible Markov processes and, hence, well defined expected payoffs according to Definition 3.4.

Definition 3.6 Let $\varepsilon > 0$ and $S \in P(N)$. An *ε-behavior* of coalition S is a (mixed) coalition behavior β_S such that $\beta_S(\pi) \in [\varepsilon, 1-\varepsilon]$ for all $\pi \in \Pi$ with $S \notin \pi$. The set of ε-behaviors of coalition S is denoted by Δ_S^ε, and the set of ε-behavior profiles by $\Delta^\varepsilon = \times_{S \in P(N)} \Delta_S^\varepsilon$. □

In an ε-behavior, every possible move is implemented with positive probability. This means that coalitions will make mistakes with some small (but positive) probability. In the theory of dynamic games the ability to account for (even one own's) possible mistakes provides one of the motivation of subgame perfection: Players specify their actions even for histories that would never be reached if they followed their strategy, and after any history their strategy needs to specify some equilibrium behavior. In this paper, we make this option of mistakes explicit as it ensures that for any two partitions π, π' there is some positive probability that a chain of coalitional moves will lead from π to π'.

Lemma 3.7 *For every $\varepsilon > 0$ and every $\beta \in \Delta^\varepsilon$ the Markov process with transition matrix P^β is irreducible.*

This lemma together with Proposition 3.2 implies that the payoff function in (4) is well defined for all $\beta \in \Delta^\varepsilon$. Observe, however, that irreducibility of the emerging Markov process is not necessary for the payoff function to be well defined: The behavior profile in Example 3.5 is not an ε-behavior, yet the payoffs are well defined.

4 Equilibrium

From here on, let $\varepsilon > 0$ and $\rho = (\rho^\pi)_{\pi \in \Pi}$ be fixed. We shall use the payoff functions in (4) to obtain an equilibrium coalition behavior that exists for all hedonic games.

4.1 Definition

Let S be a nonempty coalition and let $\pi \in \Pi$. For a given $\beta \in \Delta^\varepsilon$, we write $\beta_{-S} = (\beta_T)_{\emptyset \neq T \neq S}$ for the profile of ε-behaviors for all coalitions but S. It will also be convenient to write $\beta_S(-\pi)$ for the restriction of the behavior β_S on $\Pi \setminus \{\pi\}$. In this case, we write $(\beta_S(\pi), \beta_S(-\pi))$ for the behavior β_S. Let $\beta \in \Delta^\varepsilon$. Then $S \in P(N)$ has a profitable *one-shot deviation from β at π* if $S \notin \pi$ and there is $q \in [\varepsilon, 1-\varepsilon]$ such that

$$u_i(q, \beta_S(-\pi), \beta_{-S}) > u_i(\beta)$$

for all $i \in S$. We say that $\beta_S^*(\pi)$ is a *weak best reply* against β at π if S does not have a profitable one-shot deviation from $\left(\beta_S^*(\pi), \beta_S(-\pi), \beta_{-S}\right)$ at π. That is, a weak best reply of S against β at π takes the behavior of all coalitions and S's behavior everywhere but in π as given and specifies an optimal probability at π. We denote the set of S's weak best replies against β at π by $R_{S,\pi}^{\varepsilon}(\beta)$.[12]

Definition 4.1 A *weak ε-equilibrium* is an ε-behavior profile β such that for each $S \in P(N)$ and each $\pi \in \Pi$ it holds that $\beta_S(\pi) \in R_{S,\pi}^{\varepsilon}(\beta)$. □

That is, β is a weak ε-equilibrium if for each nonempty coalition S and each partition π the behavior β_S specifies a weak best reply $\beta_S(\pi)$ against β at π. We call such profile a "weak" equilibrium as it is only stable with respect to one-shot deviations, but not with respect to arbitrary deviations.

Example 4.2 Recall the roommate problem and the behavior profile in Example 3.1. Although this is not an ε-behavior, we have well defined payoff functions so that we can try and find weak best replies. For that purpose recall the behavior profile β in Example 3.1 and consider coalition $\{2,3\}$ at π^1. Suppose this coalition leaves π^1 with probability q. Then the corresponding transition matrix differs from P^{β} only in the second column, where p is replaced by q. The stationary distribution of the new matrix is given by $\left(0, \frac{p}{p+2q}, \frac{q}{p+2q}, \frac{q}{p+2q}, 0\right)$. So, the payoffs of players 2 and 3 are

$$u_2\left(q, \beta_{\{2,3\}}\left(-\pi^1\right), \beta_{-\{2,3\}}\right) = \frac{ap+4q}{p+2q} \quad u_3\left(q, \beta_{\{2,3\}}\left(-\pi^1\right), \beta_{-\{2,3\}}\right) = \frac{(4+a)q}{p+2q}.$$

Observe that u_3 is always increasing in q while u_2 is increasing in q for $a < 2$ and decreasing in q for $a > 2$. Thus, for $a < 2$ the only weak best response of $\{2,3\}$ at π^1 is to choose $q = 1 - \varepsilon$.[13] On the other hand, for $a \geq 2$, every $q \in [\varepsilon, 1-\varepsilon]$ is a weak best response as the interests of 2 and 3 are conflicting. □

In the previous example the set $R_{\{2,3\},\pi^1}^{\varepsilon}(\beta)$ is, depending on a, either a point set or a compact interval. This is true in general for all $\beta \in \Delta^{\varepsilon}$, $S \in P(N)$, and $\pi \in \Pi$. (See Lemma A.1 in the appendix.)

4.2 Existence

We have mentioned before that the utility function in (4) is not necessarily linear in β. The reason is as follows: Consider two behavior profiles β and γ that induce Markov processes with transition matrices P^{β} and P^{γ}, which in turn have stationary distributions μ^{β} and μ^{γ}. It can easily be verified that the convex combination $r\beta + (1-r)\gamma$ will lead to an irreducible Markov process. However, there is very little that can be said about the stationary distribution $\mu^{r\beta+(1-r)\gamma}$ of this process. In particular, it is not necessarily the case that $\mu^{r\beta+(1-r)\gamma}$ is a convex combination of μ^{β} and μ^{γ}.

This nonlinearity of the utility functions in β creates a problem as we cannot use standard arguments from normal form games to show that the set of weak best replies is convex.

[12] Note that if $S \in \pi$, then S does not have any profitable one-shot deviations at π. Hence, $R_{S,\pi}^{\varepsilon}(\beta) = \{0\}$ for all $\beta \in \Delta^{\varepsilon}$ and all $S \in P(N), \pi \in \Pi$ with $S \in \pi$.

[13] Technically, $\gamma_{1,2}$ is not a ε-behavior as $\gamma_{1,2}\left(\pi^0\right) = 1$. But it is sufficient to illustrate the concept of weak best responses without losing tractability.

Instead, we prove the following theorem that considers the stationary distribution of a convex combinations of two Markov processes whose transition matrices are identical everywhere but in one column.

Theorem 4.3 *Let X be a finite set, and let $P, Q \in [0,1]^{X\times X}$ be transition matrices of irreducible Markov processes over X, so that there is y^* with $P_{x,y} = Q_{x,y}$ for all $x \in X$ and all $y \neq y^*$. Let λ and μ be the (unique) stationary distributions of P and Q, respectively. Let $r \in [0,1]$ and define*

$$t = \frac{r\mu(y^*)}{r\mu(y^*) + (1-r)\lambda(y^*)} \tag{5}$$

Then $rP + (1-r)Q$ is the transition matrix of an irreducible Markov process, and $\nu = t\lambda + (1-t)\mu$ is the unique stationary distribution of this process.

Consider a (completely mixed) behavior profile β, and fix a partition π and a coalition $S \in P(N)$. Then, for any two strategies β_S^1 and β_S^2 that coincide with β_S everywhere but in π the transition matrices of the corresponding Markov processes differ only in column π. That is, they satisfy the condition of Theorem 4.3. Thus, we obtain the following result.

Corollary 4.4 *Let $\beta \in \Delta^\varepsilon$, $S \in P(N)$, and $\pi^* \in \Pi$ with $S \notin \pi^*$. Let $\underline{\beta}_S, \overline{\beta}_S$ be such that $\underline{\beta}_S(\pi) = \overline{\beta}_S(\pi) = \beta_S(\pi)$ for all $\pi \neq \pi^*$, and $\underline{\beta}_S(\pi^*) = \varepsilon$ and $\overline{\beta}_S(\pi^*) = 1-\varepsilon$. Let $r = \frac{1-\varepsilon-\beta_S(\pi^*)}{1-2\varepsilon}$. Then*

$$u_i(\beta) = tu_i\left(\underline{\beta}_S, \beta_{-S}\right) + (1-t)\,u_i\left(\overline{\beta}_S, \beta_{-S}\right)$$

for all $i \in N$, where t is defined as in (5).

This solves the issue outlined above: For any behavior profile β, each player's payoff from a convex combination of two deviations of S at π is a convex combination of the payoffs from the two deviations.

For all $\beta \in \Delta_S^\varepsilon$ let $R^\varepsilon(\beta) = \times_{S\in P(N)} \times_{\pi\in\Pi} R^\varepsilon_{S,\pi}(\beta)$. Then coalition behavior profile β is a weak ε-equilibrium if and only if it is a fixed point of the correspondence $\beta \mapsto R^\varepsilon(\beta)$. Thus, it is sufficient to prove that this correspondence has a fixed point. The most important part, namely convexity, follows from Corollary, 4.4. The rest of the proof is in the appendix.

Theorem 4.5 *For every hedonic coalition formation game $(N, V, \to, \rho)$ and every $\varepsilon > 0$, there is a weak ε-equilibrium.*

5 Best Responses Versus Weak Best Responses

We have seen that the definition of weak ε-equilibria ensures stability against one-shot deviation, but not necessarily against deviations at more than one state. So, the coalition formation games that we have defined in Sect. 3 lack some kind of "one-shot-principle." We shall provide an example here where a coalition does not have a one-shot deviation, i.e., is playing a weak best response, but can find a better response by changing its behavior at two states.

Let $N = \{1,2,3\}$ and v be the hedonic game given by $v(\{1\}) = 20$, $v(\{2\}) = 0$, $v(\{3\}) = 0$, $v(\{1,2\}) = (17,14)$, $v(\{1,3\}) = (17,0)$, $v(\{2,3\}) = (15,0)$ and $v(N) = (1,18,0)$. Let $\to$ be defined by the residual map γ in Example 2.3. Define three bijections $\rho_1, \rho_2, \rho_3 : \{1,\ldots,7\} \to P(N)$ by

$$(\rho_1(1), \rho_1(2), \rho_1(3), \rho_1(4), \rho_1(5), \rho_1(6), \rho_1(7)) = (\{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}),$$
$$(\rho_2(1), \rho_2(2), \rho_2(3), \rho_2(4), \rho_2(5), \rho_2(6), \rho_2(7)) = (\{1,2,3\}, \{2,3\}, \{1,3\}, \{1,2\}, \{3\}, \{2\}, \{1\}),$$
$$(\rho_3(1), \rho_3(2), \rho_3(3), \rho_3(4), \rho_3(5), \rho_3(6), \rho_3(7)) = (\{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}, \{1\}, \{2\}, \{3\}).$$

Let the partitions be numbered as in Example 2.2. Define the collection $(\rho^\pi)_{\pi\in\Pi}$ by $\rho^{\pi^0} = \rho^{\pi^3} = \rho^{\pi^4} = \rho_1$, $\rho^{\pi^1} = \rho_2$ and $\rho^{\pi^2} = \rho_3$. Let $S^0 = \{1,2\}$. We now construct a behavior profile β such that coalition S^0 has a profitable deviation at β, but no one-shot deviation. For all $T \neq S^0$ and all $\pi \in \Pi$ with $T \notin \pi$ let $\beta_T(\pi) = \frac{1}{20}$. That is, β prescribes for any $T \neq S^0$ at any π with $T \notin \pi$ to form and deviate from π to π' with probability $\frac{1}{20}$ and to remain at π with probability $\frac{19}{20}$. For each $k = 1, 2, 3, 4$ let $\beta^p_{S^0}(\pi^k) = 1 - p_k$, where $p_k \in [\varepsilon, 1-\varepsilon]$. Then the transition matrix of the Markov process associated with profile $\left(\beta^p_{S^0}, \beta_{-S^0}\right)$ is

$$P = \begin{pmatrix} p_0(1-x)^3 & x(1-x)^3(2-x) & p_2(1-x)^2x(2-x) & x(2-x) & x(x^2-3x+3) \\ 1-p_0 & (1-x)^5 & x & (1-x)^2(1-p_3) & (1-x)^3(1-p_4) \\ p_0(1-x)x & x(1-x) & p_2(1-x)^4 & p_3(1-x)^2x & p_4(1-x)^4x \\ p_0x & x(1-x)^2 & x(1-x) & p_3(1-x)^4 & p_4(1-x)^3x \\ p_0(1-x)^2x & x & (1-x)^2(1-p_2) & p_3(1-x)^3x & p_4(1-x)^5 \end{pmatrix},$$

where $x = \frac{1}{20}$. The stationary distribution of P, μ, is given by $\mu(\pi_k) = \frac{\overline{\mu}(\pi_k)}{\sum_{l=1}^k \overline{\mu}(\pi_l)}$, where

$$\begin{aligned}
\overline{\mu}(\pi_0) &= 2.659690476\ 10^{22} - 1.121619627\ 10^{22} p_2p_3p_4 + 1.611058344\ 10^{22} p_2p_3 \\
&\quad + 1.449336223\ 10^{22} p_2p_4 + 1.442311007\ 10^{22} p_3p_4 - 2.081799782\ 10^{22} p_2 \\
&\quad - 2.058328825\ 10^{22} p_3 - 1.863704548\ 10^{22} p_4 \\
\overline{\mu}(\pi_1) &= 2.621440000\ 10^{23} + 1.207584620\ 10^{23} p_0p_2p_3p_4 - 1.505504975\ 10^{23} p_0p_2p_3 \\
&\quad - 1.498726885\ 10^{23} p_0p_2p_4 - 1.479293341\ 10^{23} p_0p_3p_4 - 1.415672613\ 10^{23} p_2p_3p_4 \\
&\quad + 1.859871285\ 10^{23} p_0p_2 + 1.860700832\ 10^{23} p_0p_3 + 1.831542937\ 10^{23} p_0p_4 \\
&\quad + 1.739116855\ 10^{23} p_2p_3 + 1.748510977\ 10^{23} p_2p_4 + 1.725374942\ 10^{23} p_3p_4 \\
&\quad - 2.293812673\ 10^{23} p_0 - 2.135179264\ 10^{23} p_2 - 2.140798157\ 10^{23} p_3 \\
&\quad - 2.124770265\ 10^{23} p_4 \\
\overline{\mu}(\pi_2) &= 1.245184000\ 10^{22} - 5.434523952\ 10^{21} p_0p_3p_4 + 7.432904695\ 10^{21} p_0p_3 \\
&\quad + 7.042336316\ 10^{21} p_0p_4 + 7.023069493\ 10^{21} p_3p_4 - 9.632257219\ 10^{21} p_0 \\
&\quad - 9.608306688\ 10^{21} p_3 - 9.101201613\ 10^{21} p_4 \\
\overline{\mu}(\pi_3) &= 1.242071040\ 10^{22} - 5.307293743\ 10^{21} p_0p_2p_4 + 7.351769281\ 10^{21} p_0p_2 \\
&\quad + 6.854619389\ 10^{21} p_0p_4 + 6.950743632\ 10^{21} p_2p_4 - 9.478516665\ 10^{21} p_0 \\
&\quad - 9.634996429\ 10^{21} p_2 - 8.976693796\ 10^{21} p_4 \\
\overline{\mu}(\pi_4) &= 2.434498560\ 10^{22} - 1.319357013\ 10^{22} p_0p_2p_3 + 1.705305641\ 10^{22} p_0p_2 \\
&\quad + 1.467654760\ 10^{22} p_0p_3 + 1.695404081\ 10^{22} p_2p_3 - 1.896199018\ 10^{22} p_0 \\
&\quad - 2.191368192\ 10^{22} p_2 - 1.884302724\ 10^{22} p_3
\end{aligned}$$

The payoffs of players 1 and 2 are

$$u_1\left(\beta^p_{S^0}, \beta_{-S^0}\right) = 20(\mu_P(\pi_0) + \mu_P(\pi_2)) + 17(\mu_P(\pi_1) + \mu_P(\pi_3)) + \mu_P(\pi_4)$$
$$u_2\left(\beta^p_{S^0}, \beta_{-S^0}\right) = 14\mu_P(\pi_1) + 15\mu_P(\pi_4) + 18\mu_P(\pi_4).$$

Let $p_k^* = \frac{19}{20}$ for $k = 0, 2, 3, 4$. Then

$$\begin{aligned}
&\frac{d}{dp_0} u_1\left(\beta_{S^0}^{p^*}, \beta_{-S^0}\right) > 0 && \frac{d}{dp_0} u_2\left(\beta_{S^0}^{p^*}, \beta_{-S^0}\right) < 0\\
&\frac{d}{dp_2} u_1\left(\beta_{S^0}^{p^*}, \beta_{-S^0}\right) > 0 && \frac{d}{dp_2} u_2\left(\beta_{S^0}^{p^*}, \beta_{-S^0}\right) < 0\\
&\frac{d}{dp_3} u_1\left(\beta_{S^0}^{p^*}, \beta_{-S^0}\right) > 0 && \frac{d}{dp_3} u_2\left(\beta_{S^0}^{p^*}, \beta_{-S^0}\right) < 0\\
&\frac{d}{dp_4} u_1\left(\beta_{S^0}^{p^*}, \beta_{-S^0}\right) < 0 && \frac{d}{dp_4} u_2\left(\beta_{S^0}^{p^*}, \beta_{-S^0}\right) > 0.
\end{aligned}$$

That is, for each k any change in p_k^* makes exactly one player better off and one player worse off, so that S^0 does not have any profitable one-shot deviations from $\left(\beta_{S^0}^{p^*}, \beta_{-S^0}\right)$.

Finally, define $\hat{p}$ by $\hat{p}_0 = \hat{p}_2 = \frac{19}{20}$ and $\hat{p}_3 = \hat{p}_4 = \frac{1}{20}$. Then

$$\begin{aligned}
u_1\left(\beta_{S^0}^{\hat{p}}, \beta_{-S^0}\right) &= 17.72703770896 > 16.2479670393 = u_1\left(\beta_{S^0}^{p^*}, \beta_{-S^0}\right)\\
u_2\left(\beta_{S^0}^{\hat{p}}, \beta_{-S^0}\right) &= 9.19781147654 > 7.4664207782 = u_2\left(\beta_{S^0}^{p^*}, \beta_{-S^0}\right)
\end{aligned}$$

That is, by changing their behavior both at π^3 and at π^4 both members of S^0 can strictly improve their payoffs.

Funding Open Access funding enabled and organized by Projekt DEAL.

A Proofs

Proof of Theorem 2.5 Let π be a partition and S be a nonempty coalition. If $S \in \pi$, then $\pi' = \pi$ by **H1**, **H2** and **H3**. So, let $S \notin \pi$. Then, by **H2** and **H4**, $\pi \to_S \pi'$ only if

$$\pi' = \{\tau(i \mid \pi, S)\}_{i \in \pi(S)} \cup \{T \in \pi : T \cap S = \emptyset\}, \tag{6}$$

where τ is the residual map specified by **H4**. By **H3**, there is some π' with $\pi \to_S \pi'$. Thus, $\pi \to_S \pi'$ if and only if π' satisfies (6). In particular π' is uniquely determined by τ, which completes the proof. □

Proof of Lemma 3.7 We first show that for any two partitions $\underline{\pi}$, $\overline{\pi}$, there are an integer m, partitions $\pi^1, \ldots, \pi^m$, and coalitions $S^1, \ldots, S^{m+1}$ such that $\underline{\pi} \to_{S^1} \pi^1$, $\pi^{l-1} \to_{S^l} \pi^l$ for $l = 2, \ldots, m$, and $\pi^m \to_{S^{m+1}} \overline{\pi}$. To see this, let $\pi^* = \{\{i\}\}_{i \in N}$. It is sufficient to show that the claim is true for any $\underline{\pi}$ and $\overline{\pi} = \pi^*$, as well as for any $\overline{\pi}$ and $\underline{\pi} = \pi^*$. To see the first case, observe that for any $i, j \in N$ and any partition π with $\{j\} \in \pi$, there is π' with $\{i\} \in \pi'$ and $\pi \to_{\{i\}} \pi'$ by **H3**. Moreover, $\{i\} \in \pi'$ by **H1** and $\{j\} \in \pi'$ by **H2**. Thus, the successive deviation of singletons will lead from $\underline{\pi}$ to π^*. On the other hand, let $\overline{\pi} = \left\{R^1, \ldots, R^m\right\}$,

and let $\pi^l = \left\{R^1, \ldots, R^l\right\} \cup \{\{i\}\}_{i \in \cup_{h=l+1}^m R^h}$ for all $l = 1, \ldots, m$. Then $\pi^* \to_{R^1} \pi^1$ and $\pi^{l-1} \to_{R^l} \pi^l$ for $l = 2, \ldots, m$ by **H2**. As $\overline{\pi} = \pi^m$, the claim is proven.

Next observe that $0 < P^{\beta}_{\pi,\pi} < 1$ for all $\pi \in \Pi$ by Eq. (2). Moreover, for each partition π and each coalition S, there is a positive chance that all coalitions preceding S (according to ρ^{π}) will stay at π, so that S will be able to implement its move with positive probability. This is, in particular, true for the coalitions that have been used in the first part of the proof. Thus, for any $\pi, \pi' \in \Pi$, there is a positive chance of a move from π to π', that is, $\left(\left(P^{\beta}\right)^m\right)_{\pi',\pi} > 0$ for some $m \in \mathbb{N}$. □

Proof of Theorem 4.3 Surely, the new Markov with transition matrix $rP + (1-r)Q$ is irreducible. Thus, it has a unique stationary distribution, and it is sufficient to show that it is ν. That is, we have to show that ν is a strictly positive probability distribution over X, and that $(rP + (1-r)Q)\nu = \nu$. Clearly, $\sum_{x \in X} \nu(x) = 1$. Moreover, by construction, $r\mu(y^*) + (1-r)\lambda(y^*) > 0$, so that t is well defined. Thus, $\nu(x) > 0$ for all $x \in X$.

Further,

$$\begin{aligned}(1-r)t\lambda\left(y^*\right) - r(1-t)\mu\left(y^*\right) &= \frac{(1-r)r\mu(y^*)\lambda(y^*)}{r\mu(y^*) + (1-r)\lambda(y^*)} - \frac{r(1-r)\lambda(y^*)\mu(y^*)}{r\mu(y^*) + (1-r)\lambda(y^*)} \\ &= 0.\end{aligned}$$

Thus, recalling that $P\lambda = \lambda$ and $Q\mu = \mu$, and denoting by $P_{\cdot,y^*}$ and $Q_{\cdot,y^*}$ the y^*-th coloumn of P and Q, respectively, we find

$$\begin{aligned}(rP + (1-r)Q)\nu &= rtP\lambda + r(1-t)P\mu + (1-r)tQ\lambda + (1-r)(1-t)Q\mu \\ &= t\lambda - (1-r)tP\lambda + (1-t)\mu - r(1-t)Q\mu + r(1-t)P\mu + (1-r)tQ\lambda \\ &= t\lambda + (1-t)\mu + r(1-t)(P-Q)\mu + (1-r)t(Q-P)\lambda \\ &= \nu + r(1-t)\mu\left(y^*\right)\left(P_{\cdot,y^*} - Q_{\cdot,y^*}\right) + (1-r)t\lambda\left(y^*\right)\left(Q_{\cdot,y^*} - P_{\cdot,y^*}\right) \\ &= \nu + \left(r(1-t)\mu\left(y^*\right) - (1-r)t\lambda\left(y^*\right)\right)\left(P_{\cdot,y^*} - Q_{\cdot,y^*}\right) \\ &= \nu,\end{aligned}$$

which proves, together with Proposition 3.2, that ν is the stationary distribution of $rP + (1-r)Q$. □

Proof of Corollary 4.4 Let $\underline{P} = P^{\left(\underline{\beta}_S, \beta_{-S}\right)}$, let $\overline{P} = P^{\left(\overline{\beta}_S, \beta_{-S}\right)}$, and observe that the corresponding Markov processes are irreducible by Lemma 3.7. First note that

$$r\underline{\beta}_S\left(\pi^*\right) + (1-r)\overline{\beta}_S\left(\pi^*\right) = \frac{1-\varepsilon-\beta_S(\pi^*)}{1-2\varepsilon}\varepsilon + \frac{\beta_S(\pi^*)-\varepsilon}{1-2\varepsilon}(1-\varepsilon) = \beta_S\left(\pi^*\right),$$

so that $r\underline{\beta}_S + (1-r)\overline{\beta}_S = \beta_S$. We show that $P^{\beta} = r\underline{P} + (1-r)\overline{P}$. Surely, $P^{\beta}_{\pi',\pi} = \underline{P}_{\pi',\pi} = \overline{P}_{\pi',\pi}$ for all $\pi' \in \Pi$ and all $\pi \neq \pi^*$, so that $P^{\beta}_{\pi',\pi} = r\underline{P}_{\pi',\pi} + (1-r)\overline{P}_{\pi',\pi}$ for

all $\pi' \in \Pi$ and all $\pi \neq \pi^*$. Let l^* be such that $\rho^{\pi^*}(l^*) = S$. Then

$$\begin{aligned}
P^{\beta}_{\pi^*,\pi^*} &= \left(1 - \left(r\underline{\beta}_S(\pi^*) + (1-r)\,\overline{\beta}_S(\pi^*)\right)\right) \prod_{l \neq l^*}^{2^{|N|-1}} \left(1 - \beta_{\rho(l)}(\pi^*)\right) \\
&= r\left(1 - \underline{\beta}_S(\pi^*)\right) \prod_{l \neq l^*}^{2^{|N|-1}} \left(1 - \beta_{\rho(l)}(\pi^*)\right) \\
&\quad + (1-r)\left(1 - \overline{\beta}_S(\pi^*)\right) \prod_{l \neq l^*}^{2^{|N|-1}} \left(1 - \beta_{\rho(l)}(\pi^*)\right) \\
&= r\underline{P}_{\pi^*,\pi^*} + (1-r)\,\overline{P}_{\pi^*,\pi^*}.
\end{aligned}$$

Moreover, for $\pi' \neq \pi^*$ we have

$$\begin{aligned}
P^{\beta}_{\pi',\pi^*} &= \sum_{l:\pi^* \to_{\rho^{\pi^*}(l)} \pi'} \beta_{\rho^{\pi^*}(l)}(\pi^*) \prod_{h<l} \left(1 - \beta_{\rho^{\pi^*}(h)}(\pi^*)\right) \\
&= \sum_{l<l^*:\pi^* \to_{\rho^{\pi^*}(l)} \pi'} \beta_{\rho^{\pi^*}(l)}(\pi^*) \prod_{h<l} \left(1 - \beta_{\rho^{\pi^*}(h)}(\pi^*)\right) \\
&\quad + \left(r\underline{\beta}_{\rho^{\pi^*}(l^*)}(\pi^*) + (1-r)\,\overline{\beta}_{\rho^{\pi^*}(l^*)}(\pi^*)\right) \prod_{h<l^*} \left(1 - \beta_{\rho^{\pi^*}(h)}(\pi^*)\right) \\
&\quad + \sum_{l>l^*:\pi^* \to_{\rho^{\pi^*}(l)} \pi'} \beta_{\rho^{\pi^*}(l)}(\pi^*) \prod_{h<l, h\neq l^*} \left(1 - \beta_{\rho^{\pi^*}(h)}(\pi^*)\right) \\
&\quad \cdot \left(1 - \left(r\underline{\beta}_{\rho^{\pi^*}(l^*)}(\pi^*) + (1-r)\,\overline{\beta}_{\rho^{\pi^*}(l^*)}(\pi^*)\right)\right) \\
&= r\left(\sum_{l<l^*:\pi^* \to_{\rho^{\pi^*}(l)} \pi'} \beta_{\rho^{\pi^*}(l)}(\pi^*) \prod_{h<l} \left(1 - \beta_{\rho^{\pi^*}(h)}(\pi^*)\right)\right. \\
&\quad + \underline{\beta}_{\rho^{\pi^*}(l^*)}(\pi^*) \prod_{h<l^*} \left(1 - \beta_{\rho^{\pi^*}(h)}(\pi^*)\right) \\
&\quad \left. + \sum_{l>l^*:\pi^* \to_{\rho^{\pi^*}(l)} \pi'} \beta_{\rho^{\pi^*}(l)}(\pi^*) \left(1 - \underline{\beta}_{\rho^{\pi^*}(l^*)}(\pi^*)\right) \prod_{h<l, h\neq l^*} \left(1 - \beta_{\rho^{\pi^*}(h)}(\pi^*)\right)\right) \\
&\quad + (1-r)\left(\sum_{l<l^*:\pi^* \to_{\rho^{\pi^*}(l)} \pi'} \beta_{\rho^{\pi^*}(l)}(\pi^*) \prod_{h<l} \left(1 - \beta_{\rho^{\pi^*}(h)}(\pi^*)\right)\right. \\
&\quad + \overline{\beta}_{\rho^{\pi^*}(l^*)}(\pi^*) \prod_{h<l^*} \left(1 - \beta_{\rho^{\pi^*}(h)}(\pi^*)\right) \\
&\quad \left. + \sum_{l>l^*:\pi^* \to_{\rho^{\pi^*}(l)} \pi'} \beta_{\rho^{\pi^*}(l)}(\pi^*) \left(1 - \overline{\beta}_{\rho^{\pi^*}(l^*)}(\pi^*)\right) \prod_{h<l, h\neq l^*} \left(1 - \beta_{\rho^{\pi^*}(h)}(\pi^*)\right)\right) \\
&= r\underline{P}_{\pi',\pi^*} + (1-r)\,\overline{P}_{\pi',\pi^*}.
\end{aligned}$$

Hence, we find that $P^\beta = r\underline{P} + (1-r)\overline{P}$. Thus, by Theorem 4.3, the Markov process associated with behavior profile $\beta = \left(r\underline{\beta}_S + (1-r)\overline{\beta}_S, \beta_{-S}\right)$ has the unique stationary distribution $\nu = t\underline{\mu} + (1-t)\overline{\mu}$, where $\underline{\mu}$ and $\overline{\mu}$ are the stationary distributions of $\underline{P}$ and $\overline{P}$, respectively. Thus, by Eq. (4), we obtain

$$\begin{aligned} u_i(\beta) &= u_i\left(r\underline{\beta}_S + (1-r)\overline{\beta}_S, \beta_{-S}\right) \\ &= \sum_{\pi\in\Pi} \nu(\pi) V_i(\pi) = \sum_{\pi\in\Pi} t\lambda(\pi) V_i(\pi) + (1-t)\mu(\pi) V_i(\pi) \\ &= t u_i\left(\underline{\beta}_S, \beta_{-S}\right) + (1-t) u_i\left(\overline{\beta}_S, \beta_{-S}\right) \end{aligned}$$

for all $i \in N$. □

Proof of Theorem 4.5 It is sufficient to show that the correspondence $\Delta^\varepsilon \rightrightarrows \Delta^\varepsilon$ with $\beta \mapsto R^\varepsilon(\beta)$ has a fixed point. To that end we first show that $R^\varepsilon(\beta)$ is the product of nonempty, compact, and convex sets, and then that the correspondence is upper hemi-continuous. □

Lemma A.1 *Let $S \in P(N)$, $\beta \in \Delta^\varepsilon$, and $\pi^* \in \Pi$ with $S \notin \pi^*$. Then $R^\varepsilon_{S,\pi^*}(\beta)$ is nonempty, compact, and convex.*

Proof Let $\underline{\beta}_S$ and $\overline{\beta}_S$ be defined as in Corollary 4.4. We show that

$$R^\varepsilon_{S,\pi^*}(\beta) = \begin{cases} \{\varepsilon\} & \text{if } u_i\left(\underline{\beta}_S, \beta_{-S}\right) > u_i\left(\overline{\beta}_S, \beta_{-S}\right) \text{ for all } i \in S \\ \{1-\varepsilon\} & \text{if } u_i\left(\underline{\beta}_S, \beta_{-S}\right) < u_i\left(\overline{\beta}_S, \beta_{-S}\right) \text{ for all } i \in S \\ [\varepsilon, 1-\varepsilon] & \text{otherwise.} \end{cases}$$

The first two cases are clear, and their proofs are omitted. So, suppose that neither case applies. Then there are $i, j \in S$ (potentially $i = j$) such that

$$u_i\left(\underline{\beta}_S, \beta_{-S}\right) \le u_i\left(\overline{\beta}_S, \beta_{-S}\right) \tag{7}$$

$$u_j\left(\underline{\beta}_S, \beta_{-S}\right) \ge u_j\left(\overline{\beta}_S, \beta_{-S}\right). \tag{8}$$

Let $q \in [\varepsilon, 1-\varepsilon]$ and assume that $q \notin R^\varepsilon_{S,\pi^*}(\beta)$. Then there is $q^* \in [\varepsilon, 1-\varepsilon]$ such that $u_k\left(q^*, \beta_S(-\pi^*), \beta_{-S}\right) > u_k\left(q, \beta_S(-\pi^*), \beta_{-S}\right)$ for all $k \in S$. Let $r = \frac{1-\varepsilon-q}{1-2\varepsilon}$ and $r^* = \frac{1-\varepsilon-q^*}{1-2\varepsilon}$ and note that $(q, \beta_S(-\pi^*)) = r\underline{\beta}_S + (1-r)\overline{\beta}_S$ and, similarly, $(q^*, \beta_S(-\pi^*)) = r^*\underline{\beta}_S + (1-r^*)\overline{\beta}_S$. Define t and t^* as in (5) for r and r^*, respectively. If $t^* \ge t$, then, by Corollary 4.4 and (7),

$$\begin{aligned} u_i\left(q^*, \beta_S(-\pi^*), \beta_{-S}\right) &> u_i\left(q, \beta_S(-\pi^*), \beta_{-S}\right) = u_i\left(r\underline{\beta}_S + (1-r)\overline{\beta}_S, \beta_{-S}\right) \\ &= t u_i\left(\underline{\beta}_S, \beta_{-S}\right) + (1-t) u_i\left(\overline{\beta}_S, \beta_{-S}\right) \\ &\ge t^* u_i\left(\underline{\beta}_S, \beta_{-S}\right) + (1-t^*) u_i\left(\overline{\beta}_S, \beta_{-S}\right) \\ &= u_i\left(r^*\underline{\beta}_S + (1-r^*)\overline{\beta}_S, \beta_{-S}\right) \\ &= u_i\left(q^*, \beta_S(-\pi^*), \beta_{-S}\right) \end{aligned}$$

which is impossible; and if $t^* \leq t$, then, similarly with (8),

$$\begin{aligned} u_j\left(q^*, \beta_S\left(-\pi^*\right), \beta_{-S}\right) > u_j\left(q, \beta_S\left(-\pi^*\right), \beta_{-S}\right) &= t u_j\left(\underline{\beta}_S, \beta_{-S}\right) + (1-t)\, u_j\left(\overline{\beta}_S, \beta_{-S}\right) \\ &\geq t^* u_j\left(\underline{\beta}_S, \beta_{-S}\right) + \left(1-t^*\right) u_j\left(\overline{\beta}_S, \beta_{-S}\right) \\ &= u_j\left(q^*, \beta_S\left(-\pi^*\right), \beta_{-S}\right), \end{aligned}$$

which is impossible as well. Hence, $q \in R^{\varepsilon}_{S,\pi^*}(\beta)$, i.e., $R^{\varepsilon}_{S,\pi^*}(\beta) = [\varepsilon, 1-\varepsilon]$. □

We next show that the correspondence $\Delta^{\varepsilon} \rightrightarrows \Delta^{\varepsilon}$ with $\beta \mapsto R^{\varepsilon}(\beta)$ is upper hemicontinuous.

Lemma A.2 *The correspondence $\Delta^{\varepsilon} \rightrightarrows \Delta^{\varepsilon}$ with $\beta \mapsto R^{\varepsilon}(\beta)$ is upper hemicontinuous.*

Proof Let $(\beta^n)_{n\in\mathbb{N}}$ be a converging sequence of mixed behavior profiles $\beta^n \in \Delta^{\varepsilon}$ with $\lim_{n\to\infty} \beta^n = \beta$, and let $(\gamma^n)_{n\in\mathbb{N}}$ be a sequence with $\gamma^n \in R^{\varepsilon}(\beta^n)$ for all $n \in \mathbb{N}$. As $R^{\varepsilon}(\beta^n) \subseteq \Delta^{\varepsilon}$ for all $n \in \mathbb{N}$ and the latter is compact, there is a converging subsequence $(\gamma^{n_k})_{k\in\mathbb{N}}$ with $\gamma = \lim_{k\to\infty} \gamma^{n_k} \in \Delta^{\varepsilon}$. Assume that $\gamma \notin R^{\varepsilon}(\beta)$. Then there are $S \in P(N)$ and $\pi^* \in \Pi$ such that $\gamma_S(\pi^*) \notin R^{\varepsilon}_{S,\pi^*}(\beta)$. Thus, there is $q \in [\varepsilon, 1-\varepsilon]$, such that $u_i\left(q, \beta_S\left(-\pi^*\right), \beta_{-S}\right) > u_i\left(\gamma_S(\pi^*), \beta_S\left(-\pi^*\right), \beta_{-S}\right)$. Let $\left(q^k\right)_{k\in\mathbb{N}}$ be a sequence in $[\varepsilon, 1-\varepsilon]$ such that $\lim_{k\to\infty} q^k = q$. Moreover, let $\delta = u_i\left(q, \beta_S\left(-\pi^*\right), \beta_{-S}\right) - u_i\left(\gamma_S(\pi^*), \beta_S\left(-\pi^*\right), \beta_{-S}\right) > 0$. By the continuity of u_i there is $K^1 \in \mathbb{N}$ such that $\left|u_i\left(\gamma_S^{n_k}(\pi^*), \beta_S^{n_k}\left(-\pi^*\right), \beta_{-S}^{n_k}\right) - u_i\left(\gamma_S(\pi^*), \beta_S(\pi^*), \beta_{-S}\right)\right| < \frac{1}{2}\delta$ for all $k \geq K^1$ and all $i \in S$. For the same reason, there is $K^2 \in \mathbb{N}$ such that $\left|u_i\left(q^k, \beta_S^{n_k}\left(-\pi^*\right), \beta_{-S}^{n_k}\right) - u_i\left(q, \beta_S\left(-\pi^*\right), \beta_{-S}\right)\right| < \frac{1}{2}\delta$ for all $k \geq K^2$ and all $i \in S$. Thus, for all $k \geq \max\left\{K^1, K^2\right\}$ and all $i \in S$

$$\begin{aligned} u_i\left(q^k, \beta_S^{n_k}\left(-\pi^*\right), \beta_{-S}^{n_k}\right) &> u_i\left(q, \beta_S\left(-\pi^*\right), \beta_{-S}\right) - \frac{1}{2}\delta \\ &\geq u_i\left(\gamma_S\left(\pi^*\right), \beta_S\left(-\pi^*\right), \beta_{-S}\right) + \frac{1}{2}\delta \\ &> u_i\left(\gamma_S^{n_k}\left(\pi^*\right), \beta_S^{n_k}\left(-\pi^*\right), \beta_{-S}^{n_k}\right). \end{aligned}$$

But this is a contradiction as $\gamma^{n_k}(\pi^*) \in R^{\varepsilon}_{S,\pi^*}(\beta^{n_k})$ by construction. Hence, $\gamma \in R^*(\beta)$, which proves upper hemicontinuity. □

By Lemmas A.1, A.2, and Kakutani's fixed point theorem, there is $\beta \in \Delta^{\varepsilon}$ such that $\beta \in R^{\varepsilon}(\beta)$. By the definition of $R^{\varepsilon}(\beta)$, such a behavior profile is a weak ε-equilibrium. □

References

1. Atay A, Mauleon A, Vannetelbosch V (2021) A bargaining set for the roommate problem. J Math Econ 94:102465
2. Banerjee SHK, Sönmez T (2001) Core in a simple coalition formation game. Soc Choice Welfare 19:135–153
3. Bloch F, van den Nouweland A (2020) Farsighted stability with heterogeneous expectations. Games Econom Behav 121:32–54
4. Bogomolnaia A, Jackson M (2002) The stability of hedonic coalition structures. Games Econom Behav 38:201–230
5. Bondareva ON (1963) Some applications of linear programming methods to the theory of cooperative games. Problemy Kybernetiki 10:119–139
6. Chwe MSY (1994) Farsighted coalitional stability. J Econ Theory 63:299–325
7. Diamantoudi E, Xue L (2003) Farsighted stability in hedonic games. Soc Choice Welfare 21:39–61

8. Drèze J, Greenberg J (1980) Hedonic coalitions: optimalilty and stability. Econometrica 48:987–1003
9. Harsanyi JC (1974) An equilibrium-point interpretation of stable sets and a proposed alternative definition. Manage Sci 20:1472–1495
10. Hart S, Kurz M (1983) Endogenous formation of coalitions. Econometrica 51:1047–1064
11. Hirata D, Kasuya Y, Tomoeda K (2021) Stability against robust deviations in the roommate problem. Games Econom Behav 130:474–498
12. Iehlé V (2007) The core-partition of hedonic games. Math Soc Sci 54:176–185
13. Karos D, Robles L (2021) Full farsighted rationality. Games Econom Behav 130:409–424
14. Kimya M (2020) Equilibrium coalitional behavior. Theor Econ 15:669–714
15. Ray D, Vohra R (2015) The farsighted stable set. Econometrica 83:977–1011
16. Shapley L (1967) On balanced sets and the core. Naval Res Logist Q 14:453–460
17. Stokey NL, Lucas REJ (1999) Recursive methods in economic dynamics, 5th edn. Havard University Press