Finding Nash Stable Coalitions under Membership Rights in Boolean Hedonic Games

Ari Conati[1] , **Andreas Niskanen**[1] , **Ronald de Haan**[2] and **Matti Järvisalo**[1]
[1]University of Helsinki, Finland
[2]University of Amsterdam, the Netherlands
ari.conati@helsinki.fi, andreas.niskanen@helsinki.fi, me@ronalddehaan.eu, matti.jarvisalo@helsinki.fi

Abstract

Boolean hedonic games are a class of cooperative games involving multiple agents in which agents aim to form coalitions based on individual agents' preferences. In this work, we provide complexity results and exact algorithms for the task of forming Nash stable coalitions under different membership rights in the dichotomous setting where agents specify preferences for which coalitions they are happy/unhappy to join. The membership rights specify veto rights for coalitions, allowing a coalition to forbid an individual agent from moving (exiting the current coalition or entering another coalition) even if the agent themself would become more happy to move. We establish that various problem variants and their refinements in this setting are often situated on the second level of the polynomial hierarchy, complete for Σ_2^p. Building on the complexity results, we develop Boolean satisfiability (SAT) based counterexample-guided abstraction refinement algorithms for the Σ_2^p problem variants and empirically evaluate a first-of-kind implementation of the approaches.

1 Introduction

A coalition among agents is formed to achieve a common goal. Coalitions are central in various interactions between individuals, from economics, politics and social life (Darmann et al. 2018) to robots and software agents (Saad et al. 2011). Coalition formation, the task of forming one or more coalition from a set of agents based on the preferences of individual agents, is a key topic in the realm of multi-agent systems (Rahwan et al. 2015), and has been extensively studied from the perspective of knowledge representation and reasoning. Hedonic preferences refer to every agent caring only about which agents are in their coalition, without caring about the composition of other coalitions.

In this work we focus on coalition formation with hedonic preferences, or simply, hedonic games (Drèze and Greenberg 1980; Banerjee, Konishi, and Sönmez 2001; Bogomolnaia and Jackson 2002). An outcome of a hedonic game is a partition of the set of agents into disjoint sets, the coalitions. Various variants of hedonic games have been studied in the literature (Aziz and Savani 2016; Elkind and Rothe 2016). Two key features involving hedonic games are (i) how preferences are expressed and (ii) what constitutes a sought-after partition. In terms of (i), various types of preference representation schemes have been proposed, including e.g., ordinal preferences (Hajduková 2006), additively separable preferences (Bogomolnaia and Jackson 2002; Aziz, Brandt, and Seedig 2013), and hedonic coalition nets (Elkind and Wooldridge 2009). In this work we consider Boolean hedonic games in which preferences are dichotomous (Aziz et al. 2016; Peters 2016), i.e., are stated in terms of which coalitions an agent is happy or unhappy to belong to in terms of which other agents are in their coalition. In terms of (ii), typically hedonic games are studied with the goal of achieving outcomes that are stable under different notions such as core stability or Nash stability (Banerjee, Konishi, and Sönmez 2001; Bogomolnaia and Jackson 2002; Ballester 2004; Dimitrov et al. 2006). For intuition, a partition is Nash stable if no single agent has an incentive to move to another (possibly empty) coalition, and core stable if no group of agents would strictly prefer the group to their assigned coalition. Often different notions of stability give rise to NP-hardness for deciding if a given hedonic game admits a stable partition (Ballester 2004; Sung and Dimitrov 2007a; Sung and Dimitrov 2010; Woeginger 2013b; Woeginger 2013a). For example, for Boolean hedonic games deciding the existence of a Nash stable partition is NP-complete (Peters 2016) while for strict core stability the problem is presumably even harder, namely Σ_2^p-complete (Peters 2017).

In this work we study computational aspects of finding Nash stable partitions in Boolean hedonic games under membership rights (Sung and Dimitrov 2007b; Karakaya 2011). Membership rights refer to refinements of stability notions that have been proposed to allow for modelling and solving hedonic games in cases in which a coalition has the right to disallow single-agent deviations (Brandt, Bullinger, and Tappe 2024)—an agent from moving to another coalition or from their assigned coalition. In terms of single-agent deviations, Nash stability only requires that, keeping the coalition structure otherwise intact, an agent may move to another coalition if they individually turn from unhappy to happy by making the move. Nash stability hence considers "free entry" (the agent is allowed to enter another coalition), "free exit" (the agents is allowed to exit its current coalition) membership rights, or FE-FX in short. The more stringent requirement to free entry/exit are approved entry/exit (AE/AX), giving rise to the refined notions of *individual stability* (i.e., AE-FX), *contractual stability* (i.e., FE-AX), and *contractual individual stability* (i.e., AE-AX) in the literature (Bogomolnaia and Jackson 2002; Ballester 2004; Sung and Dimitrov 2007b; Karakaya 2011).

Our contributions come both in terms of new complexity results and exact algorithms for finding stable coalitions in Boolean hedonic games. In terms of complexity results, we establish Σ_2^p-completeness results for deciding the existence of a Nash stable partition under different membership rights combined with requiring Pareto-optimality (the requirement that no partitions exist which would make some agent happier and other agents at least as happy) or core stability. Specifically, the problems of deciding if a given Boolean hedonic game admits a partition which is (i) both Pareto optimal and Nash stable under AE-FX; (ii) both Pareto optimal and Nash stable under FE-AX; and (iii) both core stable and Nash stable under FE-AX are all Σ_2^p-complete. Furthermore, as we show, deciding the existence of a partition that is Nash stable under FE-AX is NP-complete. As most of our of complexity results suggest that the problems in question do not admit a polynomial-size encoding into Boolean satisfiability (SAT) (Biere et al. 2021), we complement the complexity results by designing SAT-based counterexample-guided abstraction refinement (CEGAR) (Clarke et al. 2003; Clarke, Gupta, and Strichman 2004) algorithms for each of the Σ_2^p-complete problems. The CEGAR algorithms are based on iteratively using two SAT solvers, one as a base NP-abstraction solver and one as a counterexample-checking NP-oracle. We furthermore empirically evaluate our open-source implementations of the CEGAR approaches and show empirically that the algorithms can scale beyond 100 agents within minutes.

2 Related Work

Notions of Nash stability, including those involving membership rights, have been studied for various types of hedonic games. Finding AE-FX Nash stable partitions has been shown to be computationally hard for various types of restricted non-dichotomous preferences (Peters and Elkind 2015). By contrast, in Boolean hedonic games an AE-FX or AE-AX Nash stable partition is guaranteed to exist and can be found in polynomial time (Peters 2016). Brandt, Bullinger, and Wilczynski (2023) showed, however, that finding a sequence of (AE) deviations leading to an AE-FX Nash stable coalition from an arbitrary partition is NP-hard even in Boolean games. Hybrid membership rights, where deviations require majority rather than unanimous approval, have also been proposed (Brandt, Bullinger, and Tappe 2022). Additionally, Aziz et al. (2015) consider various notions of welfare maximization in fractional hedonic games, in analogy to how we consider Pareto optimality in the dichotomous case.

Algorithmic approaches to hedonic games, particularly for notions involving membership rights, are somewhat scarce. Bullinger and Kraiczy (2024) develop probabilistic algorithms for finding FE-AX and AE-FX Nash stable partitions in randomized games where agents' (additively separable) preferences are drawn from a probability distribution. Bullinger et al. (2025) obtain algorithms for finding FE-AX and AE-FX Nash stable partitions in hedonic games with additive preferences, but for a variant where coalition sizes are constrained to fixed bounds, and for which such partitions are polynomial-time computable. Fichtenberger and Rey (2021) propose sublinear algorithms for determining whether a so-called graphical hedonic game (with many agents) is "nearly" stable under various notions, including membership rights based notions. Additionally, Bredereck, Elkind, and Igarashi (2019) proposed a polynomial-time algorithm for computing AE-FX partitions in hedonic diversity games (with single-peaked preferences). In hedonic diversity games, a dichotomous variant has also been proposed (Darmann 2023).

Finally, beyond hedonic games, algorithmic approaches using SAT or related techniques have been earlier proposed for other problems studied in computational social choice, including voting (Boixel and Endriss 2020; Conati, Niskanen, and Järvisalo 2024; Conati et al. 2024) and fair allocation (Bredereck et al. 2021; Bredereck et al. 2023; Conati et al. 2025).

3 Preliminaries

We begin with necessarily definitions related to propositional satisfiability, Boolean hedonic games, and Nash stability under membership rights.

3.1 Boolean Satisfiability

For a Boolean variable x there are two literals, x and $\neg x$. A clause C is a disjunction ($\vee$) of literals. A conjunctive normal form (CNF) formula F is a conjunction ($\wedge$) of clauses. For convenience we view clauses as sets of literals and CNF formulas as sets of clauses. A truth assignment τ maps each variable to 0 (false) or 1 (true), and is extended to literals via $\tau(\neg x) = 1 - \tau(x)$, to clauses via $\tau(C) = \max\{\tau(l) \mid l \in C\}$, and to formulas via $\tau(F) = \min\{\tau(C) \mid C \in F\}$. The *Boolean satisfiability* problem (SAT) asks if a given CNF formula F has an assignment τ with $\tau(F) = 1$; if so, F is satisfiable (and τ a satisfying truth assignment to F), and otherwise unsatisfiable. A SAT solver is a complete decision procedure for SAT (or, in practice, an implementation of such a decision procedure which, given enough computational resources, is guaranteed to output a satisfying truth assignment for any satisfiable CNF formula, and determine unsatisfiability for any unsatisfiable CNF formula).

Problems hard for higher levels of the polynomial hierarchy do not, under standard complexity-theoretic assumptions, admit polynomial-sized SAT encodings. A central paradigm for solving such problems is SAT-based counterexample-guided abstracton refinement (CEGAR) (Clarke et al. 2003; Clarke, Gupta, and Strichman 2004), typically instantiated with two SAT solvers for Σ_2^p-complete problems. The first solver is used to iteratively solve an *abstraction* in NP of the problem to obtain candidate solutions. The abstraction overapproximates the actual solutions: each actual solution is a solution to the abstraction, but in general not vice versa. The second solver is used to check the current candidate for *counterexamples*, witnessing the fact that the candidate is not an actual solution. In case a counterexample is found, it is used to generate *refinement* constraints which are added to the first solver to prevent the counterexample (and potentially other non-solutions) from occurring in subsequent iterations.

Search terminates once either the abstraction becomes unsatisfiable (implying that there are no actual solutions) or an actual solution is found (a candidate has no counterexample).

3.2 Hedonic Games

In a hedonic game $(A, \succeq)$, a set of *agents* $A = \{1 \ldots n\}$ have *preferences* $\succeq = (\succeq_i)_{i \in A}$, where for each $i \in A$, $\succeq_i$ is a complete, reflexive, and transitive relation over all possible *coalitions* $\mathcal{C}_i = \{C \subseteq A \mid i \in C\}$ containing agent i. A *partition* is a collection of coalitions $\pi \subseteq 2^A$ with $\bigcup_{C \in \pi} C = A$, and for every $C, D \in \pi$ with $C \neq D$ it holds that $C \cap D = \emptyset$. For a partition π, we denote the coalition to which agent i is assigned by $\pi(i) \in 2^A$.

We consider Boolean hedonic games (Aziz et al. 2016; Peters 2016), where preferences are *dichotomous*, separating coalitions into preferred and non-preferred sets of agents. Formally, a preference relation $\succeq_i$ is dichotomous if we can partition the coalitions $\mathcal{C}_i$ into two disjoint sets $\mathcal{C}_i^+$ and $\mathcal{C}_i^-$, where for any two coalitions $C, D \in \mathcal{C}_i$, it holds that $C \succeq_i D$ if and only if $C \in \mathcal{C}_i^+$ or $D \in \mathcal{C}_i^-$. We represent a dichotomous preference relation $\succeq_i$ by a propositional formula ϕ_i over variables $A \setminus \{i\}$, where ϕ_i evaluates to true under τ if and only if the assignment τ corresponds to a coalition in $\mathcal{C}_i^+$ (i.e., $\{i\} \cup \{j \in A \setminus \{i\} \mid \tau(j) = 1\} \in \mathcal{C}_i^+$).

Of interest in hedonic games are criteria of partitions of the set of agents. A key such criterion is *stability*, where we consider a class of possible *deviations* by the agents. A partition is stable for a given class of deviations if and only if no deviations are possible. We consider *single-agent deviations* where a partition is altered only with respect to the deviating agent. That is, a single-agent deviation by agent $i \in A$ transforms a partition π to π' so that $\pi(i) \neq \pi'(i)$, and for all agents $j \neq i$, it holds that $\pi(j) \setminus \{i\} = \pi'(j) \setminus \{i\}$. We denote such a deviation by $\pi \xrightarrow{i} \pi'$. A *Nash deviation* is a single-agent deviation performed by agent i with $\pi'(i) \succ_i \pi(i)$. A partition where no agent can perform a Nash deviation is *Nash stable*.

We consider Nash stability under *membership rights*, which place restrictions on the set of allowed Nash deviations by requiring consent of the coalitions the agent is entering and/or exiting. Towards a formal definition, we follow (Brandt, Bullinger, and Tappe 2024). Let $C \subseteq A$ be a coalition and $i \in A$ an agent. The *favor-in* set of C wrt. i is the set of agents in C (excluding i) who strictly favor having i inside C, i.e., $F_{\text{in}}(C, i) = \{j \in C \setminus \{i\} \mid C \cup \{i\} \succ_j C \setminus \{i\}\}$. The *favor-out* set of C wrt. i is, dually, the set of agents in C (excluding i) who strictly favor having i outside C, i.e., $F_{\text{out}}(C, i) = \{j \in C \setminus \{i\} \mid C \setminus \{i\} \succ_j C \cup \{i\}\}$.

An approved-entry (AE) Nash deviation (an individual deviation) is a Nash deviation $\pi \xrightarrow{i} \pi'$ with $F_{\text{out}}(\pi'(i), i) = \emptyset$. Analogously, an approved-exit (AX) Nash deviation (a contractual deviation) is a Nash deviation $\pi \xrightarrow{i} \pi'$ with $F_{\text{in}}(\pi(i), i) = \emptyset$. We denote free entry and free exit, i.e., Nash deviation without required approval for entering or exiting coalitions, respectively, by FE and FX. Hence Nash stability is more unambiguously denoted FE-FX Nash stability. A partition π is AE-FX Nash stable if no agent can perform an AE Nash deviation; FE-AX Nash stable if no agent can perform an AX Nash deviation; and AE-AX Nash stable if no agent can perform a Nash deviation which is both AE and AX. We again note that In the literature AE-FX, FE-AX, and AE-AX Nash stability are also called *individual stability*, *contractual stability*, and *contractual individual stability*, respectively; we use the terms synonymously. The more uniform and perhaps more descriptive *E-*X naming scheme was introduced by Karakaya (2011).

Example 1. *Suppose three agents have dichotomous preferences with $\mathcal{C}_1^+ = \{\{1, 2\}\}$, $\mathcal{C}_2^+ = \{\{2, 3\}\}$, and $\mathcal{C}_3^+ = \{\{1, 2, 3\}\}$. Consider the partition $\pi = \{\{1, 2\}, \{3\}\}$. Agent 2 has an available Nash deviation to the coalition $\{3\}$, and agent 3 has an available Nash deviation to $\{1, 2\}$ (i.e., π is not FE-FX Nash stable). The deviation of agent 2 to $\{3\}$ is an AE Nash deviation (agent 3 is indifferent to the move) but not an AX Nash deviation (agent 1 would veto). Conversely, the deviation of agent 3 to $\{1, 2\}$ is an AX Nash deviation ($\pi(3)$ is a singleton) but not an AE Nash deviation (agent 1 would veto). Thus π is not Nash stable under AE-FX or FE-AX, but it is AE-AX Nash stable, since neither Nash deviation is valid under both approved entry and exit.*

We also consider the more classical notions of *core stability* and *Pareto optimality* in combination with membership rights. Given a partition π, a core-blocking coalition is a coalition $C \notin \pi$ where for each agent $i \in C$, $C \succ_i \pi(i)$: π is core-stable if there is no coalition which blocks it. A partition π' Pareto-dominates π if (i) for each agent $i \in A$, $\pi'(i) \succeq_i \pi(i)$, and (ii) there exists some agent $j \in A$ where $\pi'(j) \succ_j \pi(j)$: we then say a π is Pareto-optimal if no partition dominates it.

4 Nash Stability under Approved Entry

We start with AE-FX Nash stability (i.e., individual Nash stability), where an agent requires approval from a coalition's members in order to join (deviate to) their coalition. We establish Σ_2^{p}-completeness for deciding the existence of an AE-FX Nash stable partition which is also Pareto-optimal and detail a SAT-based CEGAR algorithm in line with this computational complexity.

4.1 Complexity

Although in Boolean hedonic games an AE-FX Nash stable partition is guaranteed to exist and can be found in polynomial time (Peters 2016), we will show that deciding the existence of an AE-FX Nash stable partition which is also Pareto-optimal is complete for Σ_2^{p}. Proof details are provided under Appendix A for readability.

Lemma 1. *For any Boolean hedonic game G and for any agent i in G, we can modify G to a game G' involving the addition of only a constant number of new agents such that the partitions for G' that are both Pareto optimal (PO) and Nash stable under approved entry and free exit (AE-FX)—when projected to the agents appearing in G—consists of exactly those partitions for G that are both Pareto optimal (PO) and Nash stable under approved entry and free exit (AE-FX) and where agent i is satisfied.*

Lemma 2. *For any Boolean hedonic game G and for any agent i in G, we can modify G to a game G' involving the addition of only a constant number of new agents such that the partitions for G' that are both Pareto optimal (PO) and Nash stable under approved entry and free exit (AE-FX)—when projected to the agents appearing in G—consists of exactly those partitions for G that are both Pareto optimal (PO) and Nash stable under approved entry and free exit (AE-FX) and where agent i appears in a singleton coalition.*

Theorem 3. *Deciding if a given Boolean hedonic game admits a partition that is both Pareto optimal and Nash stable under approved entry and free exit (AE-FX, i.e., individually Nash stable) is Σ_2^{p}-complete.*

Proof. For membership in Σ_2^{p}, a polynomial-time (alternating) algorithm can first use existential nondeterminism to guess a partition π and check that it is Nash stable under approved entry and free exit (AE-FX), and then use universal nondeterminism to verify that there are no Pareto improvements.

To show hardness, we give a reduction from $\exists\forall$-QBF-SAT. Let $\chi = \exists X_1 \forall X_2 \psi$ be a quantified Boolean formula, where $\psi = t_1 \vee \cdots \vee t_m$ is a (quantifier-free) formula in 3DNF. Let $X_1 = \{x_1, \ldots, x_{n_1}\}$ and $X_2 = \{x_{n_1+1}, \ldots, x_{n_2}\}$.

We will construct a Boolean hedonic game as follows, that admits a partition that is both Pareto optimal (PO) and Nash stable under approved entry and free exit (AE-FX) if and only if χ is true. As the set of agents, we take $A = \{x_i, \overline{x}_i, w_i \mid 1 \le i \le n_1\} \cup \{x_j, \overline{x}_j \mid n_1+1 \le j \le n_2\} \cup \{y_k \mid 1 \le k \le m\} \cup \{z\}$.

The next step is to define the formulas expressing the agents' preferences. Before we describe these formulas, we briefly introduce some notation that we will use. For each term t_k in ψ, the formula $[\![\neg t_k]\!]$ is obtained from t_k as follows. We start with $\neg t_k$, and write it into negation normal form—i.e., as a clause. We replace each negative literal $\neg x_i$, by $\overline{x}_i$. We are now ready to define the agents' preferences:

$$\begin{aligned}
\phi_{x_i} &= w_i && \text{for } 1 \le i \le n_1\\
\phi_{\overline{x}_i} &= w_i && \text{for } 1 \le i \le n_1\\
\phi_{w_i} &= (x_i \leftrightarrow \neg\overline{x}_i) \wedge \neg z\\
\phi_{x_j} &= \neg\overline{x}_j && \text{for } n_1+1 \le j \le n_2\\
\phi_{\overline{x}_j} &= \neg x_j && \text{for } n_1+1 \le j \le n_2\\
\phi_{y_k} &= ([\![\neg t_k]\!] \wedge y_{k+1}) \vee \neg z && \text{for } 1 \le k < m\\
\phi_{y_m} &= [\![\neg t_m]\!] \vee \neg z\\
\phi_z &= y_1
\end{aligned}$$

In addition, we invoke Lemma 1 (a polynomial number of times) to enforce that all agents w_i, for $1 \le i \le n_1$, are satisfied in any partition that is both Pareto optimal (PO) and Nash stable under approved entry and free exit (AE-FX). Also, we invoke Lemma 2 (a polynomial number of times) to enforce that all agents x_j and $\overline{x}_j$, for $n_1+1 \le j \le n_2$, all agents y_k, for $1 \le k \le m$, and agent z are in a singleton coalition in any partition that is both Pareto optimal (PO) and Nash stable under approved entry and free exit (AE-FX). We defer a full proof of correctness to the appendix.

Intuitively, a solution α for χ corresponds to a partition π where agents for literals x_i over X_1 that are false in α are in a coalition with (the corresponding) w_i, and all other agents are in a singleton coalition. A counterexample β for α being a solution for χ corresponds to a Pareto improvement involving a 'grand' coalition involving the agents for literals that are true in $\alpha \cup \beta$. ☐

4.2 Algorithm

We continue by detailing a SAT-based CEGAR procedure for finding a partition which is both AE-FX Nash stable and Pareto-optimal. To this end, we first develop a SAT encoding for AE-FX Nash stable partitions, building upon a previously presented encoding for envy-freeness in hedonic games (Conati et al. 2025).

Let $G = (A, \succeq)$ be a Boolean hedonic game where $\succeq = (\succeq_i)_{i \in A}$ is dichotomous; that is, each $\succeq_i$ is represented by a propositional formula ϕ_i over variables $A \setminus \{i\}$. To represent a partition π, we instantiate Boolean variables $p_{i,j}$ for each $i, j \in A$, $i \ne j$. For a satisfying truth assignment τ, we interpret $\tau(p_{i,j}) = 1$ iff $\pi(i) = \pi(j)$. We ensure symmetry by treating $p_{i,j}$ and $p_{j,i}$ as the same variable, and transitivity via

$$\textsc{Transitivity}(A) = \bigwedge_{\substack{i,j,k \in A\\ i \ne j \ne k}} ((p_{i,j} \wedge p_{j,k}) \to p_{i,k}).$$

For each agent $i \in A$, we define

$$\phi_i^* = \phi_i[j \mapsto p_{i,j} \mid j \in A \setminus \{i\}].$$

In words, ϕ_i^* evaluates to true iff $\pi(i) \in A_i^+$. In addition, for each $i, j \in A$ with $i \ne j$, we define

$$\phi_i^*[j] = \phi_i[k \mapsto p_{j,k} \mid k \in A \setminus \{i,j\}][j \mapsto \top]$$

which evaluates to true iff $\pi(j) \in A_i^+$.

Let $L = \{i \in A \mid \{i\} \succeq_i C \text{ for all } C \subseteq A, i \in C\}$ be the set of agents who prefer to be in a singleton coalition. This set is precomputed by checking whether ϕ_i evaluates to true under the assignment $\tau(p_{i,j}) = 0$ for all $j \in A$, $j \ne i$. We encode individual rationality as $\varphi_{\mathrm{IR}}(G) = \bigwedge_{i \in L} \phi_i^*$.

Now, the formula

$$\textsc{LetEnter}(i,j) = \phi_i^* \to \phi_i^*[p_{i,j} \mapsto \top]$$

encodes that agent i approves the entry of agent j in their coalition, i.e., if agent i is happy, they remain happy with j included. Finally, AE-FX Nash stability is expressed by the formula $\varphi_{\text{AE-FX-NS}}(G)$ defined as $\varphi_{\mathrm{IR}}(G) \wedge$

$$\bigwedge_{\substack{i \in A \setminus L\\ j \in A \setminus \{i\}}} \left(\Big(\phi_i^*[j] \wedge \bigwedge_{\substack{k \in A\\ k \ne i}} (p_{j,k} \to \textsc{LetEnter}(k,i)) \Big) \to \phi_i^* \right)$$

where we interpret $p_{i,i} = \top$ for each agent $i \in A$.

Proposition 4. *Let $G = (A, \succeq)$ be a Boolean hedonic game. If G admits an AE-FX Nash stable partition π, then the assignment τ defined by $\tau(p_{i,j}) = 1$ iff $\pi(i) = \pi(j)$ for each $i, j \in A$, $i \ne j$, satisfies $\varphi_{\text{AE-FX-NS}}(G)$. Conversely, if τ is a satisfying assignment to $\varphi_{\text{AE-FX-NS}}(G)$, then the partition π formed by assigning $\pi(i) = \pi(j)$ iff $\tau(p_{i,j}) = 1$ is AE-FX Nash stable.*

Algorithm 1 SAT-based CEGAR for finding Nash stable and Pareto optimal partitions. **Input**: Stability notion $S \in \{\text{AE-FX-NS}, \text{FE-AX-NS}\}$, Boolean hedonic game $G = (A, \succeq)$ with preferences represented as propositional formulas $(\phi_i)_{i \in A}$.

1: $\varphi \leftarrow \textsc{Transitivity}(A)$
2: **while true do**
3: $\quad (result, \tau_{\text{abs}}) \leftarrow \text{SAT}(\varphi \wedge \varphi_S(G))$
4: $\quad$ **if** $result = unsat$ **then return false**
5: $\quad \pi_{\text{abs}} \leftarrow \textsc{Partition}(\tau_{\text{abs}})$
6: $\quad$ **while** $result = sat$ **do**
7: $\quad\quad \varphi_{\text{dom}} \leftarrow \bigwedge_{i \in \textsc{Happy}(\tau_{\text{abs}})} \phi_i^* \wedge \bigvee_{i \in \textsc{Unhappy}(\tau_{\text{abs}})} \phi_i^*$
8: $\quad\quad (result, \tau) \leftarrow \text{SAT}(\varphi \wedge \varphi_S(G) \wedge \varphi_{\text{dom}})$
9: $\quad\quad$ **if** $result = sat$ **then**
10: $\quad\quad\quad \tau_{\text{abs}} \leftarrow \tau, \pi_{\text{abs}} \leftarrow \textsc{Partition}(\tau)$
11: $\quad \varphi_{\text{dom}} \leftarrow \bigwedge_{i \in \textsc{Happy}(\tau_{\text{abs}})} \phi_i^* \wedge \bigvee_{i \in \textsc{Unhappy}(\tau_{\text{abs}})} \phi_i^*$
12: $\quad (result, \tau_{\text{cex}}) \leftarrow \text{SAT}(\varphi \wedge \varphi_{\text{dom}})$
13: $\quad$ **if** $result = unsat$ **then return** π_{abs}
14: $\quad \varphi \leftarrow \varphi \wedge \bigvee_{i \in \textsc{Unhappy}(\tau_{\text{cex}})} \phi_i^*$

The CEGAR procedure for finding an AE-FX Nash stable and Pareto-optimal partition is presented as Algorithm 1. The abstraction φ encodes valid partitions by enforcing transitivity (line 1). Entering the main loop of the algorithm, we iteratively call a SAT solver on the abstraction in conjunction with $\varphi_{\text{AE-FX-NS}}(G)$ encoding AE-FX Nash stability (line 3). If no solution to the abstraction exists, there is no Pareto-optimal AE-FX Nash stable partition either, and we return **false** (line 4). Otherwise, from the solution τ_{abs} we obtain a candidate partition π_{abs} by setting $\pi_{\text{abs}}(i) = \pi_{\text{abs}}(j)$ iff $\tau_{\text{abs}}(p_{i,j}) = 1$ for all $i, j \in A$, $i \neq j$ (line 5). Letting $\textsc{Happy}(\tau_{\text{abs}}) = \{i \in A \mid \tau_{\text{abs}}(\varphi_i^*) = 1\}$ and $\textsc{Unhappy}(\tau_{\text{abs}}) = A \setminus \textsc{Happy}(\tau_{\text{abs}})$, we proceed by iteratively (line 6) encoding a partition which Pareto-dominates the current solution (line 7), and asking a SAT solver for such a partition while still enforcing AE-FX Nash stability (line 8) until reaching unsatisfiability. In the positive case, we update the current partition π_{abs} (line 10). It now remains to check with one SAT solver call whether there exists a counterexample, that is, a partition which Pareto-dominates the current solution π_{abs} but is not AE-FX Nash stable (line 11–12). If there is no counterexample partition, the current solution is AE-FX Nash stable and Pareto-optimal, so we return it (line 13). Otherwise, we refine the abstraction by adding a constraint which excludes all partitions which are Pareto-dominated by the counterexample partition (line 14).

Proposition 5. *Let $G = (A, \succeq)$ be a Boolean hedonic game. With input S = AE-FX-NS and G, if G admits an AE-FX Nash stable and Pareto-optimal partition, then Algorithm 1 returns such a partition, and if no such partition exists, Algorithm 1 returns* ***false****.*

5 Nash Stability under Approved Exit

We continue by considering FE-AX Nash stability (i.e., contractual Nash stability), providing both complexity results and SAT-based CEGAR algorithms for different variants of its decision problem.

5.1 Complexity

In contrast to AE-FX Nash stability, in Boolean hedonic games FE-AX Nash stability has received less attention in the literature. Indeed, no polynomial-time algorithm for finding such a partition is known. As we will show, no FE-AX Nash stable partition is guaranteed to exist, and deciding the existence of such a partition is NP-complete. In turn, combining FE-AX Nash stability with Pareto-optimality or core stability results in Σ_2^{p}-completeness. We start with the NP-completeness result.

Lemma 6. *For any Boolean hedonic game G and for any agent i in G, we can modify G to a game G' involving the addition of only a constant number of new agents such that the partitions for G' that are Nash stable under free entry and approved exit (FE-AX)—when projected to the agents appearing in G—consists of exactly those partitions for G that are Nash stable under free entry and approved exit (FE-AX) and where agent i is satisfied.*

Proposition 7. *Given a Boolean hedonic game, a partition that is Nash stable under free entry and approved exit (FE-AX, i.e., contractually Nash stable) does not always exist. Deciding the existence of a partition that is Nash stable under free entry and approved exit for a given Boolean hedonic game is NP-complete.*

Proof. Membership in NP follows from the fact that checking Nash stability under free entry and approved exit (FE-AX) for a given partition is polynomial-time computable. To show NP-hardness, we give a reduction from 3SAT. Let $\psi = c_1 \wedge \cdots \wedge c_m$ be a propositional formula in 3CNF, over the variables $X = \{x_1, \ldots, x_n\}$.

We will construct a Boolean hedonic game as follows, that admits a partition that is Nash stable under free entry and approved exit (FE-AX) if and only if ψ is satisfiable. As the set of agents, we take $A = \{x_i, \overline{x}_i \mid 1 \leq i \leq n\} \cup \{y_j \mid 1 \leq j \leq m\}$. The next step is to define the formulas expressing the agents' preferences. Before we describe these formulas, we briefly introduce some notation that we will use. For each clause c_k in ψ, the formula $\llbracket c_k \rrbracket$ is obtained from c_k. We replace each negative literal $\neg x_i$, by $\overline{x}_i$. We are now ready to define the agents' preferences, as follows.

$$
\begin{aligned}
\phi_{x_i} &= \neg \overline{x}_i && \text{for } 1 \leq i \leq n \\
\phi_{\overline{x}_i} &= \neg x_i && \text{for } 1 \leq i \leq n \\
\phi_{y_j} &= \llbracket c_k \rrbracket \wedge y_{j+1} && \text{for } 1 \leq j < m \\
\phi_{y_m} &= \llbracket c_m \rrbracket
\end{aligned}
$$

In addition, we invoke Lemma 6 (a polynomial number of times) to enforce that all agents x_i and $\overline{x}_i$, for $1 \leq i \leq n_1$, and all agents y_j, for $1 \leq j \leq m$, are satisfied in any partition that is Nash stable under free entry and approved exit.

By the use of Lemma 6, partitions that are Nash stable under free entry and approved exit (FE-AX) are exactly those where all agents are satisfied. By definition of the agents' preferences, these correspond exactly to satisfying truth assignments for ψ. We defer a full proof of correctness to the appendix. $\square$

We now turn to establishing that combining FE-AX Nash stability with Pareto-optimality or core stability results in Σ_2^{p}-completeness.

Lemma 8. *For any Boolean hedonic game G and for any agent i in G, we can modify G to a game G' involving the addition of only a constant number of new agents such that the partitions for G' that are Nash stable under free entry and approved exit (FE-AX)—when projected to the agents appearing in G—consists of exactly those partitions for G that are Nash stable under free entry and approved exit (FE-AX) and where agent i appears in a singleton coalition.*

Theorem 9. *Deciding if a given Boolean hedonic game admits a partition that is both Pareto optimal and Nash stable under free entry and approved exit (FE-AX, i.e., contractually Nash stable) is Σ_2^{p}-complete.*

Proof. For membership in Σ_2^{p}, a polynomial-time (alternating) algorithm can first use existential nondeterminism to guess a partition π and check that it is Nash stable under free entry and approved exit (FE-AX), and then use universal nondeterminism to verify that there are no Pareto improvements.

The proof of Σ_2^{p}-hardness is similar to the proof of Theorem 3, where instead of Lemmas 1 and 2 we use Lemmas 6 and 8. Note that it seems at first sight that agent z could deviate to the coalition containing y_1, because they seem to be in their own singleton coalition, but the use of Lemma 8 rules out this deviation. ☐

Theorem 10. *Deciding if a given Boolean hedonic game admits a partition that is both core stable (CS) and Nash stable under free entry and approved exit (FE-AX, i.e., contractually Nash stable) is Σ_2^{p}-complete.*

Proof. For membership in Σ_2^{p}, a polynomial-time (alternating) algorithm can first use existential nondeterminism to guess a partition π and check that it is Nash stable under free entry and approved exit (FE-AX), and then use universal nondeterminism to verify that there are no core improvements.

To show hardness, we give a reduction from $\exists\forall$-QBF-SAT. Let $\chi = \exists X_1 \forall X_2 \psi$ be a quantified Boolean formula, where $\psi = t_1 \vee \cdots \vee t_m$ is a (quantifier-free) formula in 3DNF. Let $X_1 = \{x_1, \ldots, x_{n_1}\}$ and $X_2 = \{x_{n_1+1}, \ldots, x_{n_2}\}$.

We will construct a Boolean hedonic game as follows, that admits a partition that is both core stable (CS) and Nash stable under free entry and approved exit (FE-AX) if and only if χ is true. As the set of agents, we take $I = \{x_i, \overline{x}_i, w_i \mid 1 \leq i \leq n_1\} \cup \{x_j, \overline{x}_j \mid n_1+1 \leq j \leq n_2\} \cup \{y_k \mid 1 \leq k \leq m\} \cup \{z\}$.

The next step is to define the formulas expressing the agents' preferences. Before we describe these formulas, we briefly introduce some notation that we will use. For each term t_k in ψ, the formula $[\![\neg t_k]\!]$ is obtained from t_k as follows. We start with $\neg t_k$, and write it into negation normal form—i.e., as a clause. We replace each negative literal $\neg x_i$, by $\overline{x}_i$. We are now ready to define the agents' preferences, as follows.

$$
\begin{aligned}
\phi_{x_i} &= (z \wedge y_1) \vee w_i && \text{for } 1 \leq i \leq n_1 \\
\phi_{\overline{x}_i} &= (z \wedge y_1) \vee w_i && \text{for } 1 \leq i \leq n_1 \\
\phi_{w_i} &= (x_i \leftrightarrow \neg \overline{x}_i) \wedge \neg z && \\
\phi_{x_j} &= \neg \overline{x}_j \wedge (z \wedge y_1) && \text{for } n_1 + 1 \leq j \leq n_2 \\
\phi_{\overline{x}_j} &= \neg x_j \wedge (z \wedge y_1) && \text{for } n_1 + 1 \leq j \leq n_2
\end{aligned}
$$

$$
\begin{aligned}
\phi_{y_k} &= ([\![\neg t_k]\!] \wedge y_{k+1}) \wedge z && \text{for } 1 \leq k < m \\
\phi_{y_m} &= [\![\neg t_m]\!] \wedge z && \\
\phi_z &= y_1 &&
\end{aligned}
$$

In addition, we invoke Lemma 6 (a polynomial number of times) to enforce that all agents w_i, for $1 \leq i \leq n_1$, are satisfied in any partition that is Nash stable under free entry and approved exit (FE-AX). Also, we invoke Lemma 8 (a polynomial number of times) to enforce that all agents x_j and $\overline{x}_j$, for $n_1 + 1 \leq j \leq n_2$, all agents y_k, for $1 \leq k \leq m$, and agent z are in a singleton coalition in any partition that is Nash stable under free entry and approved exit (FE-AX). We defer a full proof of correctness to the appendix.

Intuitively, a solution α for χ corresponds to a partition π where agents for literals x_i over X_1 that are false in α are in a coalition with (the corresponding) w_i, and all other agents are in a singleton coalition. A counterexample β for α being a solution for χ corresponds to a Pareto improvement involving a 'grand' coalition involving the agents for literals that are true in $\alpha \cup \beta$. ☐

5.2 Algorithms

Towards a SAT encoding of FE-AX Nash stability, recall the SAT encoding for AE-FX Nash stability presented in Section 4, in particular the formulas ϕ_i^* for $i \in A$ and $\phi_i^*[j]$ for $i, j \in A$, $i \neq j$, as well as the set of agents L who prefer to be in a singleton coalition. We define the formula

$$\textsc{LetExit}(i, j) = \phi_i^* \rightarrow \phi_i^*[p_{i,j} \mapsto \bot]$$

which encodes that agent i approves the exit of agent j from their coalition, i.e., if agent i is happy, they remain happy with j excluded.

Instead of individual rationality, we now need to ensure that agents in L whose coalitions approve their exit must be satisfied, as otherwise they would deviate to a singleton coalition. This is encoded by

$$\psi_{\text{FE-AX-NS}}(G) = \bigwedge_{\substack{i \in L \\ j \in A \setminus \{i\}}} ((p_{i,j} \rightarrow \textsc{LetExit}(j, i)) \rightarrow \phi_i^*) .$$

For agents not in L, we ensure that there is no FE-AX Nash deviation via the formula $\chi_{\text{FE-AX-NS}}(G)$ defined as

$$\bigwedge_{\substack{i \in A \setminus L \\ j \in A \setminus \{i\}}} \left(\Big(\phi_i^*[j] \wedge \bigwedge_{\substack{k \in A \\ k \neq i,j}} (p_{i,k} \rightarrow \textsc{LetExit}(k, i)) \Big) \rightarrow \phi_i^* \right) .$$

Finally, $\varphi_{\text{FE-AX-NS}}(G) = \psi_{\text{FE-AX-NS}}(G) \wedge \chi_{\text{FE-AX-NS}}(G)$ encodes FE-AX Nash stability.

Algorithm 2 SAT-based CEGAR for finding Nash stable and core stable partitions. **Input**: Boolean hedonic game $G = (A, \succeq)$ with preferences represented as propositional formulas $(\phi_i)_{i \in A}$.

1: $\varphi \leftarrow \textsc{Transitivity}(A)$
2: **while true do**
3: $\quad (result, \tau_{\text{abs}}) \leftarrow \text{SAT}(\varphi \wedge \varphi_{\text{FE-AX-NS}}(G))$
4: $\quad$ **if** $result = unsat$ **then return false**
5: $\quad \varphi_{\text{bc}} \leftarrow \bigwedge_{i \in \textsc{Happy}(\tau_{\text{abs}})} \neg i \wedge \bigvee_{i \in \textsc{Unhappy}(\tau_{\text{abs}})} (i \rightarrow \phi_i)$
6: $\quad (result, \tau_{\text{cex}}) \leftarrow \text{SAT}(\varphi_{\text{bc}})$
7: $\quad$ **if** $result = unsat$ **then return** $\textsc{Partition}(\tau_{\text{abs}})$
8: $\quad \varphi \leftarrow \varphi \wedge \bigvee_{i \in \text{BC}(\tau_{\text{cex}})} \phi_i^*$

Proposition 11. *Let $G = (A, \succeq)$ be a Boolean hedonic game. If G admits an FE-AX Nash stable partition π, then the assignment τ defined by $\tau(p_{i,j}) = 1$ iff $\pi(i) = \pi(j)$ for each $i, j \in A$, $i \neq j$, satisfies $\varphi_{\text{FE-AX-NS}}(G)$. Conversely, if τ is a satisfying assignment to $\varphi_{\text{FE-AX-NS}}(G)$, then the partition π formed by assigning $\pi(i) = \pi(j)$ iff $\tau(p_{i,j}) = 1$ is FE-AX Nash stable.*

We note that a SAT-based CEGAR procedure detailed in Section 4 as Algorithm 1 for the combination of AE-FX Nash stability and Pareto-optimality can also be readily applied to FE-AX Nash stability and Pareto optimality.

Proposition 12. *Let $G = (A, \succeq)$ be a Boolean hedonic game. With input $S =$ FE-AX-NS and G, if G admits a FE-AX Nash stable and Pareto-optimal partition, then Algorithm 1 returns such a partition, and if no such partition exists, Algorithm 1 returns **false**.*

For core stability, the CEGAR procedure presented as Algorithm 2 proceeds as follows. Similarly to Algorithm 1, the abstraction encodes valid partitions as symmetric and transitive relations (line 1). We obtain a candidate solution by solving the abstraction under the constraint that the partition must be FE-AX Nash stable (line 3); if no such partition exists, we return **false** as no solution exists (line 4). Otherwise, we encode a blocking coalition, i.e., a subset of unhappy agents who would be happy in their own coalition, as a propositional formula (line 5), and call a SAT solver (line 6) asking for such a counterexample. If no counterexample exists, we return the candidate partition, as it is both core stable and AE-FX Nash stable (line 7). If, on the other hand, a blocking coalition $\text{BC}(\tau_{\text{cex}}) = \{i \in A \mid \tau_{\text{cex}}(i) = 1\}$ exists, we refine the abstraction by adding a constraint stating that at least one of the agents in the blocking coalition must be happy in subsequent solutions (line 8).

Proposition 13. *Let $G = (A, \succeq)$ be a Boolean hedonic game. With G as input, if G admits a FE-AX Nash stable and core stable partition, then Algorithm 2 returns such a partition, and if no such partition exists, Algorithm 2 returns **false**.*

6 Empirical Evaluation

We implemented the algorithms described in the previous sections for finding Nash stable partitions under approved entry or approved exit (for single-agent deviations) while satisfying Pareto-optimality or core stability. Note that, in contrast to the second-level complexity of identifying such partitions generally, a partition which is both core stable and AE-FX Nash stable can be identified in polynomial time (Peters 2016). The implementation makes use of PySAT (Ignatiev, Morgado, and Marques-Silva 2018), and we employ the SAT solver CaDiCaL (version 1.9.5) (Biere et al. 2024) as the core reasoning engine. The open-source implementation and experimental data are available at https://bitbucket.org/coreo-group/sat4hg/. All experiments were run on 2.40-GHz Intel Xeon Gold 6148 CPUs and 381-GB memory.

	average time (max)					
#a	PO+NS_AEFX		PO+NS_FEAX		CS+NS_FEAX	
20	0.3	(0.5)	0.3	(0.6)	0.3	(0.6)
40	2.0	(3.4)	1.5	(2.9)	1.6	(2.8)
60	11.8	(29.7)	5.6	(13.8)	5.3	(7.8)
80	34.0	(94.1)	14.1	(35.5)	14.5	(24.0)
100	95.9	(269.9)	28.8	(61.9)	29.7	(48.9)
120	234.5	(563.0)	50.9	(82.2)	52.4	(95.2)
140	503.8	(1632.2)	84.9	(142.1)	91.7	(162.1)
160	1082.4	(2789.3)	131.5	(215.2)	143.6	(271.6)

Table 1: Runtimes for deciding existence of partitions: Pareto optimal Nash stable under approved entry (PO+NS_AEFX) and approved exit (PO+NS_FEAX); core stable and Nash stable under approved exit (CS+NS_FEAX).

As benchmarks, we generated preferences for the agents semi-randomly in accordance with a social network graph, with each edge of the graph representing an acquaintanceship. The graph itself was constructed according to the Barabási-Albert model (Barabási and Albert 1999). Each edge of the graph was assigned either positive ("friendly" relationship) or negative ("antagonistic" relationship) polarity. Finally, for each agent i, a preference formula in disjunctive normal form with up to 3 terms was generated, where each term had up to 3 literals $p_{i,j}$ (for i, j linked by a positive edge) or $\neg p_{i,j}$ (for i, j linked by a negative edge). We generated 20 instances for each number of agents $20, 40, \ldots, 160$.

The results are summarized in Table 1 as the mean and maximum runtimes for finding Pareto-optimal, AE-FX Nash stable partitions (left column), Pareto-optimal, FE-AX Nash stable partitions (middle column), and core stable, FE-AX Nash stable partitions (right column) for different numbers of agents #a. We observe that, for each problem variant, our algorithms enable finding sought-after partitions rather quickly, despite theoretical complexity barriers. Notably, for instances with at most 40 agents, runtimes never exceed 5 seconds, and for instances with 100 agents, runtimes never exceed 5 minutes. Interestingly, runtimes are highest for the combination of Pareto optimality with Nash stability under approved entry, despite AE-FX Nash stability being polynomial-time computable on its own.

7 Conclusion

We studied coalition formation in the setting of Boolean hedonic games under different membership rights, specifically in problem settings for which computational complexity has so-far not been established and which are currently lacking practical exact algorithms. Our complexity results range from polynomial-time computability to completeness for NP and, most often, completeness for Σ_2^{p}. From the practical perspective, we developed SAT-based CEGAR algorithms for the Σ_2^{p}-complete problem variants and empirically evaluated our implementation of the algorithms, showing that they can scale beyond 100 agents within minutes.

Acknowledgements

Work financially supported by Research Council of Finland (grant 356046) and Finnish Doctoral Program Network in Artificial Intelligence AI-DOC (VN/3137/2024-OKM-6). The authors thank Finnish Computing Competence Infrastructure (FCCI) for computational and data storage resources.

The project is supported by the EuroHPC Joint Undertaking and its members including top-up funding by Ministry of Education and Culture. Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EuroHPC. Neither the European Union nor EuroHPC can be held responsible for them.

Funded by the European Union

EuroHPC Joint Undertaking

AI Declaration

The authors have not employed any generative AI tools.

A Proofs

Proof of Lemma 1. Take an arbitrary Boolean hedonic game G and an arbitrary agent i in G. We will describe how to modify G to obtain the game G' with the desired properties. Suppose that i's preference is given by the formula ϕ_i. We will add two new agents j and k, and the preferences of these three agents in G' are as follows:

$$\phi_i = (j \wedge \neg k) \vee \phi_i, \quad \phi_j = (k \wedge \neg i), \quad \phi_k = (i \wedge \neg j)$$

Take a partition π for G that is both Pareto optimal (PO) and Nash stable under approved entry and free exit (AE-FX) and where agent i's preference ϕ_i is satisfied. We can extend π to a partition π' for G' that is both Pareto optimal (PO) and Nash stable under approved entry and free exit (AE-FX), by assigning k to the same coalition as i and by assigning j to its own singleton coalition. In π', agents i and k are satisfied and agent j is not. Moreover, by construction, j cannot be satisfied without making k unsatisfied, so π' is Pareto optimal. Since π is Nash stable under approved entry and free exit (AE-FX), so is π'.

Conversely, take a partition π' for G' that is both Pareto optimal (PO) and Nash stable under approved entry and free exit (AE-FX). We argue that then agent i must be satisfied by the formula ϕ_i. Suppose this were not the case. Then, the only partitions where at least one of i, j and k are satisfied are those where two of these three agents are in the same coalition. By Pareto optimality, this must be the case for π'. Suppose that i is satisfied in π'—the other two cases are entirely similar. Then i and j are in the same coalition, and j is not satisfied. Moreover, j moving to k's coalition would form an approved-entry (AE) Nash deviation, contradicting Nash stability under approved entry and free exit (AE-FX). Thus agent i is satisfied by the formula ϕ_i. So restricting π' the agents in G, yielding π, gives us a partition for G with the desired properties. □

Proof of Lemma 2. Take an arbitrary Boolean hedonic game G and an arbitrary agent i in G. Let A be the set of agents in G. We will describe how to modify G to obtain the game G' with the desired properties. We will add a new agent j' for each agent $j \in A \setminus \{i\}$, and the preferences of these new agents are given by: $\phi_{j'} = i \wedge \neg j$, for each $j \in A \setminus \{i\}$. Moreover, we invoke Lemma 1 (a polynomial number of times) to enforce that each such new agent j' must be happy in any partition for G' that is both Pareto optimal and Nash stable under approved entry and free exit.

To see that the partitions for G' have the stated required properties, it suffices to observe that the new agents j' must be happy, and by definition of their preferences, the only way in which this can happen is if they are all together in a coalition with only agent i (and none of the other agents in $A \setminus \{i\}$. Thus, when projected to the agents appearing in G, agent i is in a singleton coalition. □

Proof of Theorem 3 (continued). $(\Rightarrow)$ Suppose that there exists a partition π that is both Pareto optimal (PO) and Nash stable under approved entry and free exit (AE-FX). By construction of the preferences and by our use of Lemmas 1 and 2, we know that the only coalitions of size more than 1 in π consist of w_i and exactly one of x_i and $\overline{x}_i$, for some $1 \leq i \leq n_1$, and moreover, that each w_i is in such a coalition of size 2. This means that all agents are satisfied in π, except for z and exactly one of x_i and $\overline{x}_i$ for each $1 \leq i \leq n_1$.

Consider the following truth assignment $\alpha : X_1 \to \{0, 1\}$, where $\alpha(x_i) = 1$ if and only if agent x_i is in a singleton coalition in π. Our claim is that $\psi[\alpha]$ is a valid propositional formula. Suppose, to derive a contradiction, that there is a truth assignment $\beta : X_2 \to \{0, 1\}$ that falsifies $\psi[\alpha]$, i.e., such that $\alpha \cup \beta$ makes $\neg t_k$ true for each $1 \leq k \leq m$. We will construct a Pareto improvement for π based on β (and α), as follows. The Pareto improvement consists of the following agents joining a single grand coalition C (instead of the singleton coalitions that they are in in π): agent z, the agents y_k for $1 \leq k \leq m$, the agents x_i for $1 \leq i \leq n_1$ such that $\alpha(x_i) = 1$, the agents $\overline{x}_i$ for $1 \leq i \leq n_1$ such that $\alpha(x_i) = 0$, the agents x_j for $n_1 + 1 \leq j \leq n_2$ such that $\beta(x_j) = 1$, and the agents $\overline{x}_j$ for $n_1 + 1 \leq j \leq n_2$ such that $\beta(x_j) = 0$. The agents x_i and $\overline{x}_i$ are not satisfied in C, but they were also not satisfied in π. The agents x_j and $\overline{x}_j$ are satisfied in C, and they were satisfied in π as well. The agent z is satisfied in C, but was not satisfied in π. All that remains to show is that the agents y_k are satisfied in C. Take an arbitrary agent y_k. Since $\alpha \cup \beta$ makes $\neg t_k$ true, by construction of C, we know that y_k is satisfied by the formula $[\![\neg t_k]\!]$. Thus, the coalition C gives rise to a Pareto improvement, contradicting Pareto optimality of π. Thus $\psi[\alpha]$ is a valid formula and χ is true.

($\Leftarrow$) Suppose that χ is true, i.e., that there exists a truth assignment $\alpha : X_1 \to \{0,1\}$, such that $\psi[\alpha]$ is a valid propositional formula. Consider the following partition π, where for each $1 \leq i \leq n_1$ the agent w_i is in a coalition of size 2, namely together with $\overline{x}_i$ if $\alpha(x_i) = 1$ and with x_i if $\alpha(x_i) = 0$, and where all other agents are in a singleton coalition.

We will show that π is both Pareto optimal (PO) and Nash stable under approved entry and free exit (AE-FX). Note that π satisfies all requirements enforced by the use of Lemmas 1 and 2. The only agents that are not satisfied in π are z and those among x_i and $\overline{x}_i$ that are in a singleton coalition. However, there is no coalition that they can join to improve their satisfaction without decreasing the satisfaction of some agent in the coalition. Thus π is Nash stable under approved entry and free exit (AE-FX).

What remains to show is that π is Pareto optimal. Suppose, to derive a contradiction, that there exists a Pareto improvement π'. By construction of the preferences, the unsatisfied agents x_i and $\overline{x}_i$ cannot be satisfied without decreasing the satisfaction of another agent. Thus z (the only remaining agent not satisfied in π) must be satisfied in π'. By an inductive argument, we know that z must be in a coalition in π' with all agents y_k, for $1 \leq k \leq m$. Call this coalition C. Further, without loss of generality, we may assume that C contains exactly one of x_j and $\overline{x}_j$, for all $n_1 + 1 \leq j \leq n_2$, and those agents x_i and $\overline{x}_i$, for $1 \leq i \leq n_1$, that are not satisfied in π. If they were not included in C, adding them would not decrease their satisfaction, and by monotonicity of the formulas $[\![\neg t_k]\!]$, it would also not decrease satisfaction of the other agents in C.

Consider the truth assignment $\beta : X_2 \to \{0,1\}$ where for each $n_1+1 \leq j \leq n_2$, $\beta(x_j) = 1$ if and only if $x_j \in C$. We claim that $\alpha \cup \beta$ falsifies ψ. To see this, consider t_k for an arbitrary $1 \leq k \leq m$. Because agent y_k is satisfied in π, it must also be satisfied in C (which is a coalition in the Pareto improvement π' of π). Since $z \in C$, this can only happen if y_k is satisfied by the formula $[\![\neg t_k]\!]$. By construction of C and β, this means that $\alpha \cup \beta$ falsifies t_k. Since t_k was arbitrary, $\alpha \cup \beta$ falsifies ψ. This contradicts our assumption that $\psi[\alpha]$ is a valid propositional formula. Thus π is Pareto optimal. □

Proof of Lemma 6. Take an arbitrary Boolean hedonic game G and an arbitrary agent i in G. We will describe how to modify G to obtain the game G' with the desired properties. Suppose that i's preference is given by the formula ϕ_i. We will add four new agents j_1, j_2, k_1, k_2, and the preferences of these five agents in G' are as follows:

$$\begin{aligned}
\phi_{j_1} &= (\neg j_2 \wedge (i \vee k_1 \vee k_2)) \vee (\neg j_2 \wedge \neg i \wedge \neg k_1 \wedge \neg k_2) \\
\phi_{j_2} &= (\neg j_1 \wedge (i \vee k_1 \vee k_2)) \vee (\neg j_1 \wedge \neg i \wedge \neg k_1 \wedge \neg k_2) \\
\phi_i &= ((j_1 \vee j_2) \wedge \neg k_1) \vee \phi_i \\
\phi_{k_1} &= (j_1 \vee j_2) \wedge \neg k_2 \\
\phi_{k_2} &= (j_1 \vee j_2) \wedge \neg i
\end{aligned}$$

Next, we will argue that the partitions for G' have the stated required properties.

Take a partition π for G that is Nash stable under free entry and approved exit (FE-AX) and where agent i's preference ϕ_i is satisfied. We can extend π to a partition π' for G' that is Nash stable under free entry and approved exit (FE-AX), by adding coalitions $\{j_1, k_1\}$ and $\{j_2, k_2\}$. In π', all of the newly added agents are satisfied, so Nash stability under free entry and approved exit (FE-AX) carries over from π to π'.

Conversely, take a partition π' for G' that is Nash stable under free entry and approved exit (FE-AX) and suppose, to derive a contradiction, that agent i's preference ϕ_i in G is not satisfied by π'. We will distinguish several cases: either (i) j_1 and j_2 are in the same coalition in π', or (ii) they are not. In case (i), we get a contradiction with stability, because both j_1 and j_2 want to deviate to their own singleton coalition. In case (ii), there is always at least one of i, k_1 and k_2 that wants to deviate to a coalition containing either j_1 or j_2 (and such a deviation makes no agent in their current coalition worse off), also contradicting the assumption of stability. Thus agent i's preference ϕ_i must be satisfied by π', and thus also by π. □

Proof of Proposition 7 (continued). ($\Rightarrow$) Suppose that there exists a truth assignment $\alpha : X \to \{0,1\}$ that satisfies ψ. We will construct a partition π that is Nash stable under free entry and approved exit (FE-AX), and that satisfies all agents. We let π consist of a single large coalition C, and all agents not in C are in their own singleton coalitions in π. We define $C = \{y_j \mid 1 \leq j \leq m\} \cup \{x_i \mid \alpha(x_i) = 1\} \cup \{\overline{x}_i \mid \alpha(x_i) = 0\}$. Because α satisfies each c_j, we know that all agents are satisfied in π.

($\Leftarrow$) Suppose that there exists a partition π that is Nash stable under free entry and approved exit (FE-AX). By the use of Lemma 6, we know that π satisfies all agents. Thus by an inductive argument we know that π contains a coalition C containing all agents y_j, for $1 \leq j \leq m$. Moreover, we know that C contains at most one of x_i and $\overline{x}_i$ for each $1 \leq i \leq n$. Since the formulas $[\![c_m]\!]$ are monotonic, we may assume without loss of generality that C contains exactly one of x_i and $\overline{x}_i$ for each $1 \leq i \leq n$. Consider the truth assignment $\alpha : X \to \{0,1\}$ defined by letting $\alpha(x_i) = 1$ if and only if $x_i \in C$. Because all agents y_j are satisfied in π, and by the construction of α, α satisfies ψ. □

Proof of Lemma 8. The proof is similar to the proof of Lemma 2, where instead of Lemma 1 we invoke Lemma 6. □

Proof of Theorem 10 (continued). ($\Rightarrow$) Suppose that there exists a partition π that is both both core stable (CS) and Nash stable under free entry and approved exit (FE-AX). By construction of the preferences and by our use of Lemmas 6 and 8, we know that the only coalitions of size more than 1 in π consist of w_i and exactly one of x_i and $\overline{x}_i$, for some $1 \leq i \leq n_1$, and moreover, that each w_i is in such a coalition of size 2. This means that the agents that are satisfied in π, are exactly the following: each w_i, for $1 \leq i \leq n_1$;

and those agents among x_i and $\overline{x}_i$, for $1 \leq i \leq n_1$, that are in a coalition with the corresponding w_i. In other words, all remaining agents are unhappy.

Consider the following truth assignment $\alpha : X_1 \to \{0, 1\}$, where $\alpha(x_i) = 1$ if and only if agent x_i is in a singleton coalition in π. Our claim is that $\psi[\alpha]$ is a valid propositional formula. Suppose, to derive a contradiction, that there is a truth assignment $\beta : X_2 \to \{0, 1\}$ that falsifies $\psi[\alpha]$, i.e., such that $\alpha \cup \beta$ makes $\neg t_k$ true for each $1 \leq k \leq m$. We will construct a core deviation for π based on β (and α), as follows. The deviation consists of the coalition C consisting of the following agents: agent z, the agents y_k for $1 \leq k \leq m$, the agents x_i for $1 \leq i \leq n_1$ such that $\alpha(x_i) = 1$, the agents $\overline{x}_i$ for $1 \leq i \leq n_1$ such that $\alpha(x_i) = 0$, the agents x_j for $n_1 + 1 \leq j \leq n_2$ such that $\beta(x_j) = 1$, and the agents $\overline{x}_j$ for $n_1 + 1 \leq j \leq n_2$ such that $\beta(x_j) = 0$. All agents in C are not satisfied in π. To see this, consider the agents y_k—for all other agents in C this follows straightforwardly. Take an arbitrary agent y_k. Since $\alpha \cup \beta$ makes $\neg t_k$ true, by construction of C, we know that y_k is satisfied by the formula $[\![\neg t_k]\!]$. Thus the coalition C gives rise to a core deviation, contradicting core stability of π. Thus $\psi[\alpha]$ is a valid formula, and χ is true.

($\Leftarrow$) Suppose that χ is true, i.e., that there exists a truth assignment $\alpha : X_1 \to \{0, 1\}$, such that $\psi[\alpha]$ is a valid propositional formula. Consider the following partition π, where for each $1 \leq i \leq n_1$ the agent w_i is in a coalition of size 2, namely together with $\overline{x}_i$ if $\alpha(x_i) = 1$ and with x_i if $\alpha(x_i) = 0$, and where all other agents are in a singleton coalition.

We will show that π is both core stable (CS) and Nash stable under free entry and approved exit (FE-AX). Note that π satisfies all requirements enforced by the use of Lemmas 6 and 8. The only agents that are satisfied in π, are exactly the following: each w_i, for $1 \leq i \leq n_1$; and those agents among x_i and $\overline{x}_i$, for $1 \leq i \leq n_1$, that are in a coalition with the corresponding w_i. However, there is no coalition in π that they can join to improve their satisfaction. Thus π is Nash stable under free entry and approved exit (FE-AX).

What remains to show is that π is core stable. Suppose, to derive a contradiction, that there exists a core improvement consisting of a coalition C. By construction of the preferences, and by an inductive argument, $\{z\} \cup \{y_k \mid 1 \leq k \leq m\} \subseteq C$. Further, without loss of generality, we may assume that C contains exactly one of x_j and $\overline{x}_j$, for all $n_1 + 1 \leq j \leq n_2$, and those agents x_i and $\overline{x}_i$, for $1 \leq i \leq n_1$, that are not satisfied in π. If they were not included in C, adding them would increase their satisfaction, and by monotonicity of the formulas $[\![\neg t_k]\!]$, it would also not decrease satisfaction of the other agents in C.

Consider the truth assignment $\beta : X_2 \to \{0, 1\}$ where for each $n_1 + 1 \leq j \leq n_2$, $\beta(x_j) = 1$ if and only if $x_j \in C$. We claim that $\alpha \cup \beta$ falsifies ψ. To see this, consider t_k for an arbitrary $1 \leq k \leq m$. Because C is a core deviation, agent y_k must be satisfied in C. This means that y_k is satisfied by the formula $[\![\neg t_k]\!]$. By construction of C and β, this means that $\alpha \cup \beta$ falsifies t_k. Since t_k was arbitrary, $\alpha \cup \beta$ falsifies ψ. This contradicts our assumption that $\psi[\alpha]$ is a valid propositional formula. Thus π is core stable. $\square$

References

Aziz, H., and Savani, R. 2016. Hedonic games. In *Handbook of Computational Social Choice*. Cambridge University Press. 356–376.

Aziz, H.; Gaspers, S.; Gudmundsson, J.; Mestre, J.; and Täubig, H. 2015. Welfare maximization in fractional hedonic games. In *IJCAI*, 461–467. AAAI Press.

Aziz, H.; Harrenstein, P.; Lang, J.; and Wooldridge, M. J. 2016. Boolean hedonic games. In *KR*, 166–175. AAAI Press.

Aziz, H.; Brandt, F.; and Seedig, H. G. 2013. Computing desirable partitions in additively separable hedonic games. *Artif. Intell.* 195:316–334.

Ballester, C. 2004. NP-completeness in hedonic games. *Games Econ. Behav.* 49(1):1–30.

Banerjee, S.; Konishi, H.; and Sönmez, T. 2001. Core in a simple coalition formation game. *Soc. Choice Welf.* 18(1):135–153.

Barabási, A.-L., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286(5439):509–512.

Biere, A.; Heule, M.; van Maaren, H.; and Walsh, T., eds. 2021. *Handbook of Satisfiability - Second Edition*, volume 336 of *FAIA*. IOS Press.

Biere, A.; Faller, T.; Fazekas, K.; Fleury, M.; Froleyks, N.; and Pollitt, F. 2024. CaDiCaL 2.0. In *CAV*, volume 14681 of *LNCS*, 133–152. Springer.

Bogomolnaia, A., and Jackson, M. O. 2002. The stability of hedonic coalition structures. *Games Econ. Behav.* 38(2):201–230.

Boixel, A., and Endriss, U. 2020. Automated justification of collective decisions via constraint solving. In *AAMAS*, 168–176. IFAAMAS.

Brandt, F.; Bullinger, M.; and Tappe, L. 2022. Single-agent dynamics in additively separable hedonic games. In *AAAI*, 4867–4874. AAAI Press.

Brandt, F.; Bullinger, M.; and Tappe, L. 2024. Stability based on single-agent deviations in additively separable hedonic games. *Artif. Intell.* 334:104160.

Brandt, F.; Bullinger, M.; and Wilczynski, A. 2023. Reaching individually stable coalition structures. *ACM Trans. Economics and Comput.* 11:4:1–4:65.

Bredereck, R.; Figiel, A.; Kaczmarczyk, A.; Knop, D.; and Niedermeier, R. 2021. High-multiplicity fair allocation made more practical. In *AAMAS*, 260–268. ACM.

Bredereck, R.; Kaczmarczyk, A.; Knop, D.; and Niedermeier, R. 2023. High-multiplicity fair allocation using parametric integer linear programming. In *ECAI*, volume 372 of *FAIA*, 303–310. IOS Press.

Bredereck, R.; Elkind, E.; and Igarashi, A. 2019. Hedonic diversity games. In *AAMAS*, 565–573. IFAAMAS.

Bullinger, M., and Kraiczy, S. 2024. Stability in random hedonic games. In *EC*, 212. ACM.

Bullinger, M.; Dunajski, A.; Elkind, E.; and Gilboa, M. 2025. Single-deviation stability in additively separable hedonic games with constrained coalition sizes. *CoRR* abs/2510.12641.

Clarke, E. M.; Grumberg, O.; Jha, S.; Lu, Y.; and Veith, H. 2003. Counterexample-guided abstraction refinement for symbolic model checking. *J. ACM* 50(5):752–794.

Clarke, E. M.; Gupta, A.; and Strichman, O. 2004. SAT-based counterexample-guided abstraction refinement. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* 23(7):1113–1123.

Conati, A.; Niskanen, A.; de Haan, R.; and Järvisalo, M. 2024. Complexity results and algorithms for manipulation and bribery in judgment aggregation. In *ECAI*, volume 392 of *FAIA*, 3597–3604. IOS Press.

Conati, A.; Niskanen, A.; de Haan, R.; and Järvisalo, M. 2025. Computing efficient and envy-free allocations under dichotomous preferences using SAT. In *AAMAS*, 510–518. IFAAMAS / ACM.

Conati, A.; Niskanen, A.; and Järvisalo, M. 2024. Declarative approaches to outcome determination in judgment aggregation. *J. Artif. Intell. Res.* 81:793–836.

Darmann, A.; Elkind, E.; Kurz, S.; Lang, J.; Schauer, J.; and Woeginger, G. J. 2018. Group activity selection problem with approval preferences. *Int. J. Game Theory* 47(3):767–796.

Darmann, A. 2023. Stability and welfare in (dichotomous) hedonic diversity games. *Theory Comput. Syst.* 67(6):1133–1155.

Dimitrov, D.; Borm, P.; Hendrickx, R.; and Sung, S. C. 2006. Simple priorities and core stability in hedonic games. *Soc. Choice Welf.* 26(2):421–433.

Drèze, J. H., and Greenberg, J. 1980. Hedonic coalitions: Optimality and stability. *Econometrica* 48(4):987–1003.

Elkind, E., and Rothe, J. 2016. Cooperative game theory. In *Economics and Computation, An Introduction to Algorithmic Game Theory, Computational Social Choice, and Fair Division*, Springer texts in business and economics. Springer. 135–193.

Elkind, E., and Wooldridge, M. J. 2009. Hedonic coalition nets. In *AAMAS*, 417–424. IFAAMAS.

Fichtenberger, H., and Rey, A. 2021. Testing stability properties in graphical hedonic games. *Auton. Agents Multi Agent Syst.* 35(2):26.

Hajduková, J. 2006. Coalition formation games: a survey. *IGTR* 8(4):613–641.

Ignatiev, A.; Morgado, A.; and Marques-Silva, J. 2018. PySAT: A Python toolkit for prototyping with SAT oracles. In *SAT*, volume 10929 of *LNCS*, 428–437. Springer.

Karakaya, M. 2011. Hedonic coalition formation games: A new stability notion. *Math. Soc. Sci.* 61(3):157–165.

Peters, D., and Elkind, E. 2015. Simple causes of complexity in hedonic games. In *IJCAI*, 617–623. AAAI Press.

Peters, D. 2016. Complexity of hedonic games with dichotomous preferences. In *AAAI*, 579–585. AAAI Press.

Peters, D. 2017. Precise complexity of the core in dichotomous and additive hedonic games. In *ADT*, volume 10576 of *LNCS*, 214–227. Springer.

Rahwan, T.; Michalak, T. P.; Wooldridge, M. J.; and Jennings, N. R. 2015. Coalition structure generation: A survey. *Artif. Intell.* 229:139–174.

Saad, W.; Han, Z.; Basar, T.; Debbah, M.; and Hjørungnes, A. 2011. Hedonic coalition formation for distributed task allocation among wireless agents. *IEEE Trans. Mob. Comput.* 10(9):1327–1344.

Sung, S. C., and Dimitrov, D. 2007a. On core membership testing for hedonic coalition formation games. *Oper. Res. Lett.* 35(2):155–158.

Sung, S. C., and Dimitrov, D. 2007b. On myopic stability concepts for hedonic games. *Theory Decis.* 62:31–45.

Sung, S. C., and Dimitrov, D. 2010. Computational complexity in additive hedonic games. *Eur. J. Oper. Res.* 203(3):635–639.

Woeginger, G. J. 2013a. Core stability in hedonic coalition formation. In *SOFSEM*, volume 7741 of *LNCS*, 33–50. Springer.

Woeginger, G. J. 2013b. A hardness result for core stability in additive hedonic games. *Math. Soc. Sci.* 65(2):101–104.