Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Settling the Complexity of Popularity in Additively Separable and Fractional Hedonic Games

Martin Bullinger, **Matan Gilboa**
University of Oxford
martin.bullinger@cs.ox.ac.uk, matan.gilboa@cs.ox.ac.uk

Abstract

We study coalition formation in the framework of hedonic games. These games model the problem of partitioning a set of agents having a preference order over the coalitions they can be part of. A partition is called popular if it does not lose a majority vote among the agents against any other partition. Unfortunately, hedonic games need not admit popular partitions. We go further and settle the complexity of the existence problem concerning popularity in additively separable and fractional hedonic games by showing that it is Σ_2^p-complete in both cases. We are thus the first work that proves a completeness result of popularity for the second level of the polynomial hierarchy.

1 Introduction

We consider the task of partitioning a set of agents, say humans or machines, into disjoint coalitions. Agents have preferences regarding the coalition they are part of and a reasonable partition should reflect these preferences. This task is commonly studied in the framework of *coalition formation* and is an intriguing object of study at the intersection of economics and computer science. The typical economic setting is the formation of teams, such as working groups or political parties, but applications also consider reaching international agreements, establishing research collaboration, or forming customs unions [Ray, 2007]. Partitioning problems are also studied in other fields, such as clustering in machine learning and community detection in social science [Cohen-Addad *et al.*, 2022; Newman, 2004].

The output of a coalition formation scenario is usually measured by means of solution concepts, that capture the ideas of stability and optimality. While stability conceptualizes the prospect of agents staying in their own coalition rather than performing deviations to join other coalitions, optimality aims at global guarantees, for instance, with respect to notions of welfare. We consider the notion of *popularity*, a solution concept due to Gärdenfors [1975] that incorporates both ideas [Brandt and Bullinger, 2022].[1] Informally speaking, an outcome is popular if no other outcome wins a vote against this outcome. In social choice theory, this corresponds to the well-established notion of weak Condorcet winners [Condorcet, 1785], but popularity can be defined in any context where agents have preferences over outcomes. Popularity hence captures the idea of a status quo that cannot be defeated in a head-to-head election. For instance, students engaged in a popular research collaboration would not be able to propose a different outcome preferred by a majority.

[1] In his original work, Gärdenfors [1975] calls popular outcomes "majority assignments."

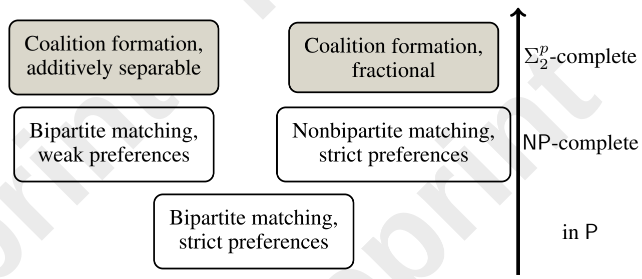

Figure 1: Complexity hierarchy of popularity in coalitional scenarios. Gray boxes refer to our main results.

Gärdenfors [1975] was the first to consider popularity in a coalitional setting. He considered bipartite matching instances and showed that stable matchings—in the sense of Gale and Shapley [1962]—are popular if the agents' preferences are strict. Interestingly, relaxing either assumption (bipartiteness or strict preferences) may lead to instances in which popular outcomes do not exist, and the corresponding decision problems become NP-complete [Biró *et al.*, 2010; Faenza *et al.*, 2019; Gupta *et al.*, 2021]. Notably, membership in NP is not trivial for this problem because one has to certify that a matching does not lose a vote against any other matching, of which there are exponentially many. This task can, however, be performed by transforming the verification of popular matchings to a maximum weight matching problem [Biró *et al.*, 2010] or a linear program that can simultaneously handle weak and incomplete preferences as well as nonbipartite instances [Kavitha *et al.*, 2011; Brandt and Bullinger, 2022].

When allowing coalitions to be of size greater than 2, we reach the typical domain of coalition formation. We consider the prominent classes of additively separable and fractional hedonic games [Bogomolnaia and Jackson, 2002; Aziz *et al.*, 2019]. Specifically, we study the following decision problems.

> ASHG-EXISTS-POPULAR (FHG-EXISTS-POPULAR)
> **Input:** Additively separable hedonic game (fractional hedonic game)
> **Question:** Does the given game admit a popular partition?

For these, we prove two sweeping hardness results. These complete the characterization of the complexity hierarchy of popularity in coalitional scenarios, as detailed in Figure 1.

Theorem 1. ASHG-EXISTS-POPULAR *is* Σ_2^p*-complete.*

Theorem 2. FHG-EXISTS-POPULAR *is* Σ_2^p*-complete, even if valuations are nonnegative.*

Our results highlight the significant computational hardness presented by popularity in coalition formation. While NP-hard problems can often be addressed in practice using SAT or ILP solvers, Σ_2^p-completeness indicates a higher level of complexity that surpasses these typical approaches.

Notably, the definition of popularity, i.e., the *existence* of an outcome such that *for all* other outcomes, a vote is not lost, suggests membership in the complexity class Σ_2^p. Still, we are the first to prove a corresponding completeness result. By contrast, previous work only establishes hardness for the first level of the polynomial hierarchy [Aziz *et al.*, 2013; Brandt and Bullinger, 2022; Cseh and Peters, 2022] or considers the simpler to analyze verification problem [Kerkmann *et al.*, 2020].

Note that the nonnegativity assumption in Theorem 2 is a strong additional restriction, which is not possible for Theorem 1. Indeed, additively separable hedonic games define agents' utilities for coalitions based on the sum of valuations of its members. Hence, forming the grand coalition containing all agents is optimal for all agents if valuations are nonnegative. By contrast, the sum of valuations is divided by the size of the coalition in fractional hedonic games, which leads to nontrivial preferences, even for nonnegative valuations.

2 Related Work

Coalition formation in the framework of hedonic games was first considered by Drèze and Greenberg [1980] and further conceptualized by Bogomolnaia and Jackson [2002], Banerjee *et al.* [2001], and Cechlárová and Romero-Medina [2001]. The book chapters by Aziz and Savani [2016] and Bullinger *et al.* [2024] provide an introduction to hedonic games.

In a general model of hedonic games, agents have to rank an exponentially large set of possible coalitions. Since this causes computational issues, a wide range of succinct preference representations has been proposed in the literature. Often, this is based on restricting attention to important meta-information about a coalition such as its size [Bogomolnaia and Jackson, 2002] or its best or worst member [Cechlárová and Romero-Medina, 2001]. Another way is to aggregate cardinal valuation functions of single agents to a utility for a coalition. We consider models that follow this latter approach, namely additively separable hedonic games (ASHGs) and fractional hedonic games (FHGs) [Bogomolnaia and Jackson, 2002; Aziz *et al.*, 2019].

Similar to the landscape of game classes, there exists a variety of solution concepts for hedonic games. We focus our discussion on the large body of research on ASHGs and FHGs. Much of this literature concerns stability, i.e., the absence of beneficial deviations to join other or form new coalitions. A common theme is that stability is usually only satisfiable for restricted domains of games, and various computational hardness results have been observed. Interestingly, there is a difference in complexity dependent on whether single agents or groups of agents perform a deviation. Whether a single agent can perform a deviation can usually be checked in polynomial time and we obtain NP-completeness results [Sung and Dimitrov, 2010; Aziz *et al.*, 2013; Brandl *et al.*, 2015; Brandt *et al.*, 2023; Brandt *et al.*, 2024]. By contrast, whether a group deviation exists is itself NP-complete to check and hence the existence of group stability, e.g., whether there exist partitions in the *core*, becomes Σ_2^p-complete [Woeginger, 2013; Peters, 2017; Aziz *et al.*, 2019]. Prior to our complexity results on popularity, these were the only problems known to be Σ_2^p-complete for hedonic games.

It is possible to achieve more positive results regarding the existence of stable outcomes by considering restricted domains [Bogomolnaia and Jackson, 2002; Dimitrov *et al.*, 2006], weakened solution concepts [Fanelli *et al.*, 2021], stability under randomized deviations [Fioravanti *et al.*, 2023], or in random games [Bullinger and Kraiczy, 2024]. For instance, symmetric utilities lead to the existence of single-deviation stability in ASHGs [Bogomolnaia and Jackson, 2002], but the same is not true in FHGs [Brandt *et al.*, 2023], and even in ASHGs, computation is still PLS-hard [Gairing and Savani, 2019].

Popularity, our main solution concept, has received less attention. Most related to our work is the paper by Brandt and Bullinger [2022] who prove NP-hardness and coNP-hardness of the existence problem for ASHGs and FHGs. In addition, they show coNP-completeness of the verification of popular partitions, a problem that was also considered by Aziz *et al.* [2013] for ASHGs.[2] Our results improve upon these results by showing Σ_2^p-completeness of the existence problem, which settles the precise complexity of popularity in ASHGs and FHGs.

Popularity has also been considered in further classes of hedonic games. Brandt and Bullinger [2022] and Cseh and Peters [2022] study it for games with coalitions bounded in size by three, and Kerkmann *et al.* [2020] consider a preference model based on the distinction of friends, enemies, and neutrals. Moreover, Kerkmann and Rothe [2020] consider popularity for a nonhedonic class of coalition formation games aimed at modeling altruism. All of these papers show coNP-completeness of the verification problem. However, while Brandt and Bullinger [2022] and Cseh and Peters [2022] ...

[2] Aziz *et al.* [2013] also consider the existence problem of popularity for ASHGs, but their proof was pointed out to be incomplete by Brandt and Bullinger [2022].

ters [2022] at least show NP-hardness, the complexity of the existence problem remains unresolved in all of these models.

3 Preliminaries

In this section, we provide the preliminaries for our work. We start with defining hedonic games, then define important solution concepts, and finally discuss the computational aspects of these solution concepts.

3.1 Succinct Classes of Cardinal Hedonic Games

Let N be a set of agents. A *coalition* is a nonempty subset of N. A coalition of size one is called a *singleton* coalition. Denote by $\mathcal{N}_i = \{S \subseteq N : i \in S\}$ the set of all coalitions agent i belongs to. A *coalition structure*, or a *partition*, is a partition π of N into coalitions. For an agent $i \in N$, we denote by $\pi(i)$ the coalition i belongs to in π.

A *hedonic game* is a pair $(N, \succsim)$, where $\succsim = (\succsim_i)_{i\in N}$ is a preference profile specifying the preferences of each agent i as a complete and transitive preference order $\succsim_i$ over $\mathcal{N}_i$. In hedonic games, agents are only concerned with the members of their own coalition which is also reflected in their preference order. Therefore, we can naturally define an associated preference order over partitions by $\pi \succsim_i \pi'$ if and only if $\pi(i) \succsim_i \pi'(i)$. For coalitions $S, S' \in \mathcal{N}_i$, we say that agent i *weakly prefers* S over S' if $S \succsim_i S'$. Moreover, we say that i *prefers* S over S' if $S \succ_i S'$. We use the same terminology for preferences over partitions.

In this paper, we assume agents rank coalitions (and by extension, partitions) by underlying utility functions $u = (u_i : \mathcal{N}_i \to \mathbb{R})_{i\in N}$. These induce a hedonic game $(N, \succsim)$ where, for every agent $i \in N$ and two coalitions $S, S' \in \mathcal{N}_i$, we define $S \succsim_i S'$ if and only if $u_i(S) \geq u_i(S')$. Hence, i prefers S over S' if and only if $u_i(S) > u_i(S')$. We say that $u_i(S)$ is i's utility for coalition S and extend this to utilities for partitions by setting $u_i(\pi) = u_i(\pi(i))$. A hedonic game together with its utility-based representation is called a *cardinal hedonic game* and is specified by the pair (N, u).

Hedonic games as introduced so far need every agent to specify a preference order or cardinal values for an exponentially large set of coalitions. By contrast, we focus on succinctly representable sub-classes of cardinal hedonic games, where the utilities are induced by the aggregation of values that each agent assigns to other members of her coalition. These games are specified by a pair (N, v), where $v = (v_i : N \to \mathbb{R})_{i\in N}$ is a vector of *valuation functions*. The quantity $v_i(j)$ denotes the value agent i assigns to agent j.

Following Bogomolnaia and Jackson [2002], an *additively separable hedonic game* (ASHG) given by the pair (N, v) is the cardinal hedonic game (N, u) where

$$u_i(S) = \sum_{j\in S\setminus\{i\}} v_i(j).$$

Hence, the utility $u_i(S)$ of agent i for coalition $S \in \mathcal{N}_i$ is defined as the sum of the values agent i assigns to the other members of her coalition.

Following Aziz *et al.* [2019], a *fractional hedonic game* (FHG) given by the pair (N, v) is the cardinal hedonic game (N, u) where

$$u_i(S) = \frac{\sum_{j\in S\setminus\{i\}} v_i(j)}{|S|}.$$

Hence, the utility $u_i(S)$ of agent i for coalition $S \in \mathcal{N}_i$ is defined as the sum of the values agent i assigns to the other members of her coalition divided by the coalition size. This quantity can be interpreted as the average value that i assigns to the members of her coalition if we include a value of 0 for herself.

3.2 Popular Partitions

We now move towards defining popularity, our main solution concept, for a given hedonic game $(N, \succsim)$. Let π and π' be two partitions of N. We denote the set of agents who prefer π over π' by $N(\pi, \pi')$, i.e., $N(\pi, \pi') = \{i \in N : \pi \succ_i \pi'\}$. For any subset of agents $M \subseteq N$, we define the *popularity margin* on M with respect to the ordered pair (π, π') to be $\phi_M(\pi, \pi') = |N(\pi, \pi') \cap M| - |N(\pi', \pi) \cap M|$. Note that in this definition, agents who are indifferent between the two partitions do not contribute to any of the two terms. When M is a singleton containing a single agent a, we use the abbreviated notation $\phi_a(\pi, \pi') = \phi_{\{a\}}(\pi, \pi')$. The definition of popularity margins is useful as sometimes it is convenient to consider restricted subsets of agents separately. Further, considering the entire set of agents, we define the *popularity margin* of the ordered pair (π, π') as $\phi(\pi, \pi') = \phi_N(\pi, \pi')$. Note that the popularity margin is antisymmetric, i.e., $\phi(\pi, \pi') = -\phi(\pi', \pi)$. We say that π is *more popular* than π' if $\phi(\pi, \pi') > 0$. Moreover, π is called *popular* if there exists no partition π' that is more popular than π, i.e., for any partition π' it holds that $\phi(\pi, \pi') \geq 0$.

Another useful concept in the context of popularity is Pareto optimality. We say that π' is a *Pareto improvement* from π if all agents weakly prefer π' over π, and at least one agent strictly prefers π' over π. If there exists no Pareto improvement from π, we say π is *Pareto-optimal*. Clearly, popular partitions are Pareto-optimal. Indeed, every Pareto improvement is a more popular partition. By contrast, Pareto-optimal partitions need not be popular. In addition, a useful observation is that it suffices to restrict attention to Pareto-optimal partitions when considering popularity [Brandt and Bullinger, 2022].

Proposition 1 (Brandt and Bullinger [2022], Proposition 4)**.** *A partition π is popular if and only if for all Pareto-optimal partitions π' it holds that $\phi(\pi, \pi') \geq 0$.*

As a consequence, whenever we postulate a more popular partition than a given partition, we may assume without loss of generality that this partition is Pareto-optimal.

3.3 Complexity Theory

We assume familiarity of the reader with basic notions of complexity theory such as polynomial-time reductions or the classes P (*deterministic polynomial time*) and NP (*nondeterministic polynomial time*). Here, we focus on the complexity class Σ_2^p in the second level of the polynomial hierarchy, which captures the problems considered in this paper. We refer to the textbooks by Papadimitriou [1994] and Arora and

Barak [2009] for an introduction to complexity and a deeper coverage of Σ_2^p.

The class Σ_2^p contains all problems Q for which there exists a polynomial-time Turing machine M and a polynomial q such that x is a Yes-instance of Q if and only if there exists a $y \in \{0,1\}^{q(|x|)}$ such that for all $z \in \{0,1\}^{q(|x|)}$ it holds that $M(x,y,z) = \text{TRUE}$. Informally speaking, this captures problems in which the solutions y of an instance x are challenged by any possible adversary z. The class is thus described by the concatenation of an existential and a universal quantifier. It therefore contains NP, which is defined by just an existential quantifier (because we can ignore the universal quantifier), and coNP, which is defined by just a universal quantifier (because we can ignore the existential quantifier). As with other complexity classes, a problem Q is said to be Σ_2^p-*hard* if for every problem in Σ_2^p, there exists a polynomial-time reduction from this problem to Q. A problem is said to be Σ_2^p-*complete* if it is Σ_2^p-hard and contained in Σ_2^p.

As a first example, we define the problem 2-QUANTIFIED 3-DNF-SAT, which is a canonical SAT problem for Σ_2^p. It is the source problem of our reductions in Theorems 1 and 2.

> 2-QUANTIFIED 3-DNF-SAT
> **Input:** Two sets $\mathcal{X} = \{x_1, \dots, x_n\}$ and $\mathcal{Y} = \{y_1, \dots, y_n\}$ of Boolean variables and a Boolean formula $\psi(\mathcal{X}, \mathcal{Y})$ over $\mathcal{X} \cup \mathcal{Y}$ in disjunctive normal form, where each of the conjunctive clauses consists of exactly three distinct literals.
> **Question:** Does there exist a truth assignment $\tau_\mathcal{X}$ to $x_1, \dots, x_n$ such that for all truth assignments $\tau_\mathcal{Y}$ to $y_1, \dots, y_n$ it holds that $\psi(\tau_\mathcal{X}, \tau_\mathcal{Y}) = \text{TRUE}$?

2-QUANTIFIED 3-DNF-SAT is exactly in the spirit of Σ_2^p. Yes-instances are described by the existence of a certificate (the truth assignment to $x_1, \dots, x_n$) such that the output of the formula is TRUE regardless of the truth assignment to $y_1, \dots, y_n$. Even more, 2-QUANTIFIED 3-DNF-SAT was shown to be Σ_2^p-complete by Stockmeyer [1977].

As a second example, we argue that ASHG-EXISTS-POPULAR and FHG-EXISTS-POPULAR are contained in Σ_2^p, as remarked by Brandt and Bullinger [2022]: One can consider a polynomial-time Turing machine with three inputs that are a hedonic game (say, an ASHG or FHG) and two partitions π and π' and it outputs TRUE if and only if $\phi(\pi, \pi') \geq 0$ in the given hedonic game. This Turing machine attests membership in Σ_2^p of the existence problem of popularity. In our proofs, we will therefore only consider hardness.

4 Popularity in ASHGs

In this section, we discuss the proof of Theorem 1. We start by describing our reduction from 2-QUANTIFIED 3-DNF-SAT. Then, in the subsequent two sections, we give an overview of the proof that satisfiability of the source instance implies the existence of a popular partition and vice versa. We focus on the key arguments and an illustration while the detailed proof is available in the full version of our paper [Bullinger and Gilboa, 2024].

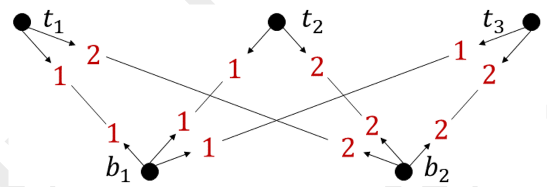

Figure 2: A No-instance of ASHG-EXISTS-POPULAR. Omitted edges imply value $-\infty$.

4.1 Setup of the Reduction

We now describe the construction of the reduction. First, we introduce the following No-instance of ASHG-EXISTS-POPULAR, on which the reduction relies; it resembles the No-instance described by **?** [**?**]Example 4]ABS11c. Suppose we have five agents, consisting of three *top* agents t_1, t_2, and t_3, and two *bottom* agents b_1 and b_2. For each $i \in \{1,2,3\}$, t_i assigns value 1 to b_1 and value 2 to b_2. Moreover, b_1 assigns value 1 to each top agent and b_2 assigns value 2 to each top agent. All other values are set to $-\infty$ (representing some sufficiently large negative value, e.g., -7 suffices here) between the agents. This instance is depicted in Figure 2.

One may verify that there exists no popular partition in this instance: It is easy to see that it is more popular to dissolve any coalition of size at least three into singletons. Hence, the interesting case is a partition of the type $\{\{t_1, b_1\}, \{t_2, b_2\}, \{t_3\}\}$, which, however, is less popular than $\{\{t_1, b_2\}, \{t_3, b_1\}, \{t_2\}\}$.

In our reduction, we construct a game which has a similar structure to this No-instance (with some additional agents). However, each top agent t_i is replaced by a set of multiple agents who, intuitively, together function in a similar way as the single agent t_i. Hence, familiarity with the above No-instance is helpful to understand the reduction as well: when a satisfying assignment to the 2-QUANTIFIED 3-DNF-SAT instance does not exist, the reduced game simulates a behaviour similar to that of this No-instance.

We proceed by describing our reduction. Suppose that we are given an arbitrary instance $(\mathcal{X}, \mathcal{Y}, \psi)$ of 2-QUANTIFIED 3-DNF-SAT. Denote by $\mathcal{C}$ the set of clauses in ψ and let $m = |\mathcal{C}|$; without loss of generality, we may assume that $m \geq 2$. We construct the following ASHG consisting of $12n+4m+1$ agents, depicted in Figure 3.

- For every variable $x \in \mathcal{X}$:
  - We create two X-agents a_x and $a_{\neg x}$, where the former represents the variable and the latter its negation. We will use α to denote any literal over $\mathcal{X}$, meaning a_α can correspond to a variable or its negation; accordingly, $a_{\neg\alpha}$ will simply correspond to the negated literal, e.g., if $\alpha = \neg x$, then $a_{\neg\alpha} = a_x$. If a_x and $a_{\neg x}$ originate in the same variable, they are called *complementary* agents.
  - We create a corresponding X_t-agent and a corresponding X_f-agent, denoted x_t and x_f, respectively. The subscripts of these agents indicate "true" and "false" and these agents are used to deduct the satisfying truth assignments from popular partitions (and vice versa).

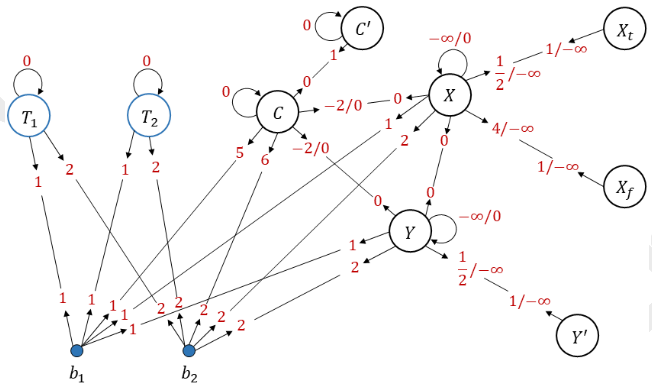

Figure 3: The reduction for the proof of Theorem 1. Omitted edges imply value $-\infty$. When two values v_1/v_2 appear, v_1 refers to corresponding agents, and v_2 to noncorresponding. Left-side agents are marked in blue. b_1 and b_2 are single agents, while the rest represent sets of agents.

- For every variable $y \in \mathcal{Y}$:
  - We create two Y-agents a_y and $a_{\neg y}$, where the former represents the variable and the latter its negation. We will use β to denote any literal over $\mathcal{Y}$, meaning a_β can correspond to a variable or its negation; $a_{\neg\beta}$ will refer to the agent corresponding to the negated literal. If a_β and $a_{\neg\beta}$ originate in the same variable, they are called *complementary* agents.
  - We create a Y'-agent a'_y corresponding to a_y, and a Y'-agent $a'_{\neg y}$ corresponding to $a_{\neg y}$ (we emphasize that, in contrast to the X-agents, which have corresponding agents as a pair, a_y and $a_{\neg y}$ each have separate Y'-agents).
- For every clause $c \in \mathcal{C}$, we create a C-agent a_c. For a literal α over $\mathcal{X}$ (or β over $\mathcal{Y}$) occurring in c, we refer to the X-agent a_α (or Y-agent a_β) as corresponding to clause c.
- We create $m-1$ agents, called C'-agents.
- We create $2n+m$ agents, called T_1 agents and another $2n+m$ agents, called T_2 agents.
- We create a single agent denoted b_1, and a single agent denoted b_2.

For each agent type, the set of all agents from that type is denoted by the name of the type (e.g., T_1 is the set of all T_1-agents). We use the terms *real agents* to refer to the X-, Y-, and C-agents, and *structure agents* to refer to all other agents. In addition, we speak of *left-side agents* to refer to b_1, b_2, and the T_1- and T_2-agents, and *right-side agents* to refer to the other agents (this terminology is based on the visualization in Figure 3). We denote by L and R the sets of all left-side and right-side agents, respectively.

We refer to Figure 3 for an overview of the valuation functions and to the full version of our paper for a detailed description. Valuations missing from the figure (as well as some of the depicted ones) correspond to a large negative constant which we indicate by a value of $-\infty$. For the reduction to work, one can, for instance, set $\infty = 6(12n+4m+1)$. This completes the description of the reduction.

When the input 2-QUANTIFIED 3-DNF-SAT instance is a No-instance, the reduced ASHG mimics the No-instance in Figure 2, where b_1 and b_2 are still single agents, but t_1 and t_2 are replaced by the sets T_1 and T_2, and the real agents correspond to the agent t_3. The real agents also encode the source instance of 2-QUANTIFIED 3-DNF-SAT, as they are representatives of the literals and clauses. The right-side structure agents provide options for "good" coalitions for the real agents.

In essence, a popular partition can only exist if all right-side agents are in good coalition which will allow for a partition corresponding to the partition $\{\{t_1, b_1\}, \{t_2, b_2\}, \{t_3\}\}$ to be popular. The good coalitions for the real agents are:

- coalitions of the type $\{a_x, x_t\}$ and $\{a_{\neg x}, x_f\}$ or $\{a_{\neg x}, x_t\}$ and $\{a_x, x_f\}$ for the X-agents,
- coalitions $\{a_\beta, a'_\beta\}$ for the Y-agents, and
- coalition $C \cup C'$ for the C-agents.

The crucial part is to determine the exact coalitions of the X-agents. Whether we form $\{a_x, x_t\}$ or $\{a_{\neg x}, x_t\}$ corresponds to a truth assignment to the $\mathcal{X}$ variables.

To prove Theorem 1, we will show that the logical formula is satisfiable if and only if there exists a popular partition in the constructed ASHG. If we have a truth assignment, we can define a partition as described above and prove that it is popular. Conversely, a popular partition has to be a structure similar to the partition described above and we can use it to extract a truth assignment. The two directions of the proof will be sketched in Sections 4.2 and 4.3.

4.2 Satisfiability Implies Popular Partition

Throughout this section, we assume that $(\mathcal{X}, \mathcal{Y}, \psi)$ is a Yes-instance of 2-QUANTIFIED 3-DNF-SAT. Hence, there is a truth assignment $\tau_\mathcal{X}$ to the variables in $\mathcal{X}$ such that for all truth assignments $\tau_\mathcal{Y}$ to the variables in $\mathcal{Y}$ it holds that $\psi(\tau_\mathcal{X}, \tau_\mathcal{Y}) = \text{TRUE}$. Consider the following partition of the agents, denoted by π^*.

- For each $x \in \mathcal{X}$, if x is assigned TRUE by $\tau_\mathcal{X}$ then $\{\{a_x, x_t\}, \{a_{\neg x}, x_f\}\} \subseteq \pi^*$, and if x is assigned FALSE by $\tau_\mathcal{X}$ then $\{\{a_{\neg x}, x_t\}, \{a_x, x_f\}\} \subseteq \pi^*$.
- Each Y-agent a_β forms a coalition with her corresponding Y'-agent a'_β.
- The coalition $C \cup C'$ is formed.
- Coalitions $T_1 \cup \{b_1\}$ and $T_2 \cup \{b_2\}$ are formed.

Our goal is to show that π^* is a popular partition. We formally prove this statement in the full version of our paper. Here, we focus on outlining key steps. Assume towards contradiction that there exists a partition π that is more popular than π^*. We wish to use π to extract a truth assignment $\tau'_\mathcal{Y}$ to the variables in $\mathcal{Y}$ such that $\psi(\tau_\mathcal{X}, \tau'_\mathcal{Y}) = \text{FALSE}$. For this, we will determine various structural insights about the partition π and finally use the coalition $\pi(b_1)$ to find both the assignment $\tau'_\mathcal{Y}$ as well as a proof that it can be used to evaluate ψ as false.

For determining the structure of π, it is good to first consider the popularity margin for certain groups of agents. By using that π^* is a very good partition for the Y'-, X_f-, X_t-, and C'-agents, we obtain the following facts

- For each $a_\beta \in Y$, it holds that $\phi_{\{a_\beta, a'_\beta\}}(\pi^*, \pi) \geq 0$.
- For each $x \in \mathcal{X}$, it holds that $\phi_{\{a_x, a_{\neg x}, x_t, x_f\}}(\pi^*, \pi) \geq 0$.
- If for every $a_c \in C$ we have that $\phi_{a_c}(\pi, \pi^*) > 0$, then $\phi_{C \cup C'}(\pi^*, \pi) = -1$. Otherwise, $\phi_{C \cup C'}(\pi^*, \pi) \geq 0$.

Together, the worst-case popularity margin of the right-side agents is thus $\phi_R(\pi^*, \pi) \geq -1$.

As a next step, we consider coalitions of left-side agents and show that

1. Agents in T_1 and T_2 cannot be in the same coalition.
2. The coalition of b_1 contains a right-side agent.
3. The coalition of b_2 does not contain a right-side agent.

Item 1 holds because these agents only gain positive value from b_1 and b_2, whereas valuations between agents in T_1 and T_2 are $-\infty$. This insight can then be leveraged to show that at least one agent of $\{b_1, b_2\}$ has to contain a right-side agent. Otherwise, it is easy to deduce that the left-side agents have a popularity margin of $\phi_L(\pi^*, \pi) \geq 0$, and furthermore no C-agent can gain positive utility in π, and therefore also $\phi_R(\pi^*, \pi) \geq 0$. Together, these two facts imply that π was not more popular. Items 2 and 3 follow with little effort from this conclusion.

We can now show that $\phi_{T_1 \cup T_2 \cup \{b_2\}}(\pi^*, \pi) \geq 0$. Together with our other insights about the popularity margins, π can only be more popular than π^* if $\phi_{b_1}(\pi^*, \pi) \leq 0$.

Next, it is easy to see that each agent in T_1 or T_2 that forms a coalition with a right-side agent would have to be in the coalition with b_1. However, by carefully analyzing $\pi(b_1)$, it can then be shown that it cannot contain agents in T_1 and T_2.

To summarize our knowledge about left-side agents, we know that b_1 forms a coalition with right-side agents only, whereas all other left-side agents form coalitions with other left-side agents.

The next step is to analyze the exact coalition of b_1 in π. It can be shown that $\pi(b_1)$ can only contain real agents (recall that $\phi_{b_1}(\pi^*, \pi) \leq 0$) and that it has to contain exactly n X-agents corresponding to the agents forming coalitions with the X_t-agents in π^*, all C-agents, and either a_β or $a_{\neg\beta}$ for every $\mathcal{Y}$ variable.

We can now extract a truth assignment $\tau'_{\mathcal{Y}}$ to $\mathcal{Y}$ from the Y-agents contained in $\pi(b_1)$. The only way that π is more popular than π^* is when all C-agents prefer π over π^* which, due to the valuations by the C-agents of the agents corresponding to their respective literals, can only happen if $\tau_{\mathcal{X}}$ and $\tau'_{\mathcal{Y}}$ evaluate every clause to FALSE. This implies that $\psi(\tau_{\mathcal{X}}, \tau'_{\mathcal{Y}}) =$ FALSE, a contradiction. We thus conclude this part of the proof.

4.3 Popular Partition Implies Satisfiability

Throughout this section, we assume that there is a popular partition π^* in the reduced ASHG. We will prove that this implies that the source instance is a Yes-instance to 2-QUANTIFIED 3-DNF-SAT. The detailed proof of this statement can be found in the full version of our paper. In this section, we give an overview of the proof.

Our main goal is to show that π^* has a structure similar to that of the popular partition defined in Section 4.2 (up to symmetries), which will enable us to extract a satisfying truth assignment to the variables in $\mathcal{X}$ by looking at the coalitions of the X_t-agents.

As a first step, we show that left-side and right-side agents cannot form a joint coalition. Suppose a coalition $S \in \pi^*$ contains both a left-side and a right-side agent. The only agents who may have a nonnegative utility in such a coalition are b_1, b_2, and real agents, and thus S must contain some combination of agents b_1 and b_2. If both b_1 and b_2 are in S, then the partition obtained from π^* by extracting b_1 and b_2 from S, and forming the coalitions $\{b_1\} \cup T_1$ and $\{b_2\} \cup T_2$, can be shown to be more popular. So, only one of b_1 and b_2 may reside in S. Denote $b_j \in S$, and $b_i \notin S$, where $i, j \in \{1, 2\}$. Hence, it is easy to see that we must have either $\pi^*(b_i) = \{b_i\} \cup T_1$ or $\pi^*(b_i) = \{b_i\} \cup T_2$. Without loss of generality, assume $\pi^*(b_i) = \{b_i\} \cup T_1$. Now, intuitively, we can think of T_1, T_2, and $S \setminus \{b_j\}$ as the agents t_1, t_2, and t_3 from the No-instance discussed in Section 4.1, respectively. A deviation analogous to that discussed in the context of this No-instance shows that this partition is not popular.

Having established that the left and right side are separated, the only possibility for π^* to be popular is if agents form coalitions with their corresponding agents, who give them positive utility. Specifically, the following must hold.

1. For the left side, we have that $\{b_1\} \cup T_1 \in \pi^*$ and $\{b_2\} \cup T_2 \in \pi^*$, or $\{b_2\} \cup T_1 \in \pi^*$ and $\{b_1\} \cup T_2 \in \pi^*$.
2. We have that $C \cup C' \in \pi^*$.
3. For each $a_\beta \in Y$, we have that $\{a_\beta, a'_\beta\} \in \pi^*$.
4. For each $x \in \mathcal{X}$, we have that $\{a_x, x_t\} \in \pi^*$ and $\{a_{\neg x}, x_f\} \in \pi^*$, or $\{a_x, x_f\} \in \pi^*$ and $\{a_{\neg x}, x_t\} \in \pi^*$.

This allows us to define the following truth assignment $\tau_{\mathcal{X}}$ to the $\mathcal{X}$ variables. For each $x \in \mathcal{X}$, x is assigned TRUE if and only if $\pi^*(a_x) = \{a_x, x_t\}$ (by Item 4, this is a valid assignment). We claim that $\tau_{\mathcal{X}}$ is a satisfying assignment to the 2-QUANTIFIED 3-DNF-SAT instance, i.e., that $\psi(\tau_{\mathcal{X}}, \tau_{\mathcal{Y}}) =$ TRUE for all truth assignments $\tau_{\mathcal{Y}}$ to the $\mathcal{Y}$ variables.

Assume otherwise, namely that there exists a truth assignment $\tau'_{\mathcal{Y}}$ to the $\mathcal{Y}$ variables such that $\psi(\tau_{\mathcal{X}}, \tau'_{\mathcal{Y}}) =$ FALSE. We will now find a partition that is more popular than π^*. Recalling Item 1, let us assume without loss of generality that $\{\{b_1\} \cup T_1, \{b_2\} \cup T_2\} \subseteq \pi^*$. Consider the partition π obtained from π^* as follows.

- Extract all $a_\alpha \in X$ such that $\{a_\alpha, x_t\} \in \pi^*$, for some $x_t \in X_t$, all $a_\beta \in Y$ such that the literal represented by a_β is assigned TRUE by $\tau'_{\mathcal{Y}}$, and all C-agents and agent b_1. With them, form a new coalition S.
- Extract b_2 from her coalition, and set $\pi(b_2) = \{b_2\} \cup T_1$.

Note that the new coalition S consists of $2n + m + 1$ agents. Moreover, by definition of $\tau_{\mathcal{X}}$, if $\tau_{\mathcal{X}}$ assigns TRUE to x, then

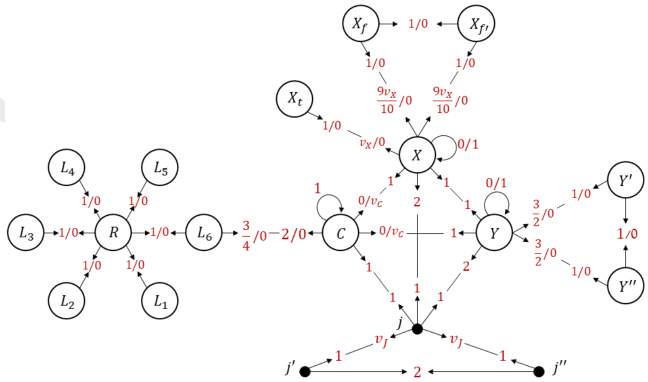

Figure 4: The reduction for the proof of Theorem 2. Each node refers to a certain agent type, i.e., to the respective set of agents. Edges indicate valuations between all agents in the respective sets. When two values v_1/v_2 appear, v_1 refers to corresponding agents, and v_2 to noncorresponding ones. Omitted edges imply value 0.

S contains a_x and if $\tau_\mathcal{X}$ assigns FALSE to x, then S contains $a_{\neg x}$. In addition, for $y \in \mathcal{Y}$, S contains a_y if $\tau'_\mathcal{Y}$ assigns TRUE to y and S contains $a_{\neg y}$ if $\tau'_\mathcal{Y}$ assigns FALSE to y.

We compute the popularity margin between π and π^*. Let $c \in \mathcal{C}$. Since $\psi(\tau_\mathcal{X}, \tau'_\mathcal{Y}) = \text{FALSE}$, we have that c has at most two literals in S assigned TRUE by $\tau_\mathcal{X}$ and $\tau'_\mathcal{Y}$. Hence, since the X- and Y-agents in S correspond to the literals assigned TRUE by $\tau_\mathcal{X}$ and $\tau'_\mathcal{Y}$, there are at most two X- or Y-agents in S to whom a_c assigns value -2 (to the other X- or Y-agents she assigns 0). Therefore, all C-agents prefer π over π^*. Thus, it is simple to check that $\phi_R(\pi^*, \pi) = -1$ (which stems from the fact that $|C'| - |C| = -1$). Furthermore, we have $\phi_L(\pi^*, \pi) = 0$ (T_1-agents prefer π, T_2-agents prefer π^*, and b_1 and b_2 are indifferent between the partitions). Altogether, we conclude that $\phi(\pi^*, \pi) = -1$, in contradiction to π^* being a popular partition. Hence, $(\mathcal{X}, \mathcal{Y}, \psi)$ is a Yes-instance of 2-QUANTIFIED 3-DNF-SAT.

5 Popularity in Nonnegative FHGs

In this section, we discuss the reduction of Theorem 2. Due to space constraints, we only describe the proof using the illustration of Figure 4. The detailed proof is available in the full version of our paper [Bullinger and Gilboa, 2024].

The reduction for ASHGs in Theorem 1 used one blown-up No-instance, where the whole combinatorics of the source problem had replaced the role of a single agent. For FHGs, the combinatorics of the source instance is still encoded similarly. Literals are still represented by X- and Y-agents and once again, they have good options to form coalitions together with their corresponding structure agents by forming coalitions $\{a_x, x_t\}$ and $\{a_{\neg x}, x_f, x_{f'}\}$ or $\{a_{\neg x}, x_t\}$ and $\{a_x, x_f, x_{f'}\}$ for the X-agents, and coalitions $\{a_\beta, a'_\beta, a''_\beta\}$ for the Y-agents. As in the proof for ASHGs, the coalitions of the X_t-agents in popular partitions correspond to satisfying truth assignments for the variables in $\mathcal{X}$.

However, the combinatorics of the source instance is not embedded in a single No-instance, but multiple No-instances are used as gadgets. The gadgets are FHGs induced by a star graph, which are known to not admit popular partitions [Brandt and Bullinger, 2022] (see the left part of Figure 4).

This implies that popular partitions have to contain coalitions of some agent in the gadgets with another agent outside the gadgets. We have one star gadget for every clause and the only agent that is linked to the gadget by a positive valuation is its corresponding clause agent. We will see that the only possibility to achieve popularity is to use this single connection by forming coalitions of the type $\{a_c, \ell_c^6\}$.

To prove Theorem 2, we will show that the logical formula is satisfiable if and only if there exists a popular partition in the constructed FHG. Every truth assignment lets us define a popular partition along these ideas, while we can show that every popular partition has such a structure and lets us extract a truth assignment.

6 Conclusion

We considered the complexity of deciding whether popular partitions exist in typical classes of hedonic games. By showing that this problem is Σ_2^p-complete, we pinpoint its precise complexity for ASHGs and FHGs with nonnegative valuation functions. Hence, allowing coalitions of size at least three can raise the complexity of popularity from completeness for the first to the second level of the polynomial hierarchy.

Our work is an important step in understanding popularity in coalition formation. However, there are still various dimensions along which a deeper understanding would be welcome. Firstly, our methods might aid in resolving the exact complexity of popularity in other classes of coalition formation games for which popularity was considered before [Kerkmann *et al.*, 2020; Kerkmann and Rothe, 2020; Brandt and Bullinger, 2022; Cseh and Peters, 2022]. Secondly, it would be interesting to consider popularity in other classes of hedonic games, such as modified fractional hedonic games [Olsen, 2012] and anonymous hedonic games [Bogomolnaia and Jackson, 2002].

Finally, popularity has the closely related concepts of strong popularity, where a partition has to strictly win the vote in a pairwise comparison against every other partition, and mixed popularity, which considers probability distributions of partitions that are popular in expectation. In the domain of matching, strong popularity and mixed popularity were first considered by Gärdenfors [1975] and Kavitha *et al.* [2011], respectively, and studied by Brandt and Bullinger [2022] in ASHGs and FHGs. In these classes, Brandt and Bullinger [2022] show that deciding whether a strongly popular partition exists is coNP-hard[3] while computing a mixed popular partition is NP-hard. The exact complexity of both problems remains open. Notably, Σ_2^p seems not to be the right complexity class for these problems because mixed popular partitions always exist and strongly popular partitions are unique whenever they exist. Resolving their complexity could lead to an intriguing complexity picture of concepts of popularity in hedonic games.

[3]This aligns with known complexity results for strong popularity in other classes of coalition formation games [Kerkmann *et al.*, 2020; Kerkmann and Rothe, 2020].

Acknowledgments

Martin Bullinger was supported by the AI Programme of The Alan Turing Institute. Matan Gilboa was supported by an Oxford-Reuben Foundation Graduate Scholarship. We would like to thank Edith Elkind for many fruitful discussions.

References

[Arora and Barak, 2009] Sanjeev Arora and Boaz Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009.

[Aziz and Savani, 2016] Haris Aziz and Rahul Savani. Hedonic games. In Felix Brandt, Vincent Conitzer, Ulle Endriss, J. Lang, and Ariel D. Procaccia, editors, *Handbook of Computational Social Choice*, chapter 15. Cambridge University Press, 2016.

[Aziz *et al.*, 2013] Haris Aziz, Felix Brandt, and Hans Georg Seedig. Computing desirable partitions in additively separable hedonic games. *Artificial Intelligence*, 195:316–334, 2013.

[Aziz *et al.*, 2019] Haris Aziz, Florian Brandl, Felix Brandt, Paul Harrenstein, Martin Olsen, and Dominik Peters. Fractional hedonic games. *ACM Transactions on Economics and Computation*, 7(2):1–29, 2019.

[Banerjee *et al.*, 2001] Suryapratim Banerjee, Hideo Konishi, and Tayfun Sönmez. Core in a simple coalition formation game. *Social Choice and Welfare*, 18:135–153, 2001.

[Biró *et al.*, 2010] Péter Biró, Robert W. Irving, and David F. Manlove. Popular matchings in the marriage and roommates problems. In *Proceedings of the 7th Italian Conference on Algorithms and Complexity (CIAC)*, pages 97–108, 2010.

[Bogomolnaia and Jackson, 2002] Anna Bogomolnaia and Matthew O. Jackson. The stability of hedonic coalition structures. *Games and Economic Behavior*, 38(2):201–230, 2002.

[Brandl *et al.*, 2015] Florian Brandl, Felix Brandt, and Martin Strobel. Fractional hedonic games: Individual and group stability. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1219–1227, 2015.

[Brandt and Bullinger, 2022] Felix Brandt and Martin Bullinger. Finding and recognizing popular coalition structures. *Journal of Artificial Intelligence Research*, 74:569–626, 2022.

[Brandt *et al.*, 2023] Felix Brandt, Martin Bullinger, and Anaëlle Wilczynski. Reaching individually stable coalition structures. *ACM Transactions on Economics and Computation*, 11(1–2):4:1–65, 2023.

[Brandt *et al.*, 2024] Felix Brandt, Martin Bullinger, and Leo Tappe. Stability based on single-agent deviations in additively separable hedonic games. *Artificial Intelligence*, 334, 2024.

[Bullinger and Gilboa, 2024] Martin Bullinger and Matan Gilboa. Settling the complexity of popularity in additively separable and fractional hedonic games. Technical report, https://arxiv.org/abs/2411.05713, 2024.

[Bullinger and Kraiczy, 2024] Martin Bullinger and Sonja Kraiczy. Stability in random hedonic games. In *Proceedings of the 25th ACM Conference on Economics and Computation (ACM-EC)*, 2024.

[Bullinger *et al.*, 2024] Martin Bullinger, Edith Elkind, and Jörg Rothe. Cooperative game theory. In Jörg Rothe, editor, *Economics and Computation: An Introduction to Algorithmic Game Theory, Computational Social Choice, and Fair Division*, chapter 3, pages 139–229. Springer, 2024.

[Cechlárová and Romero-Medina, 2001] Katarína Cechlárová and Antonio Romero-Medina. Stability in coalition formation games. *International Journal of Game Theory*, 29:487–494, 2001.

[Cohen-Addad *et al.*, 2022] Vincent Cohen-Addad, Silvio Lattanzi, Andreas Maggiori, and Nikos Parotsidis. Online and consistent correlation clustering. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 4157–4179, 2022.

[Condorcet, 1785] Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, 1785. Facsimile published in 1972 by Chelsea Publishing Company, New York.

[Cseh and Peters, 2022] Ágnes Cseh and Jannik Peters. Three-dimensional popular matching with cyclic preferences. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 309–317, 2022.

[Dimitrov *et al.*, 2006] Dinko Dimitrov, Peter Borm, Ruud Hendrickx, and Shao C. Sung. Simple priorities and core stability in hedonic games. *Social Choice and Welfare*, 26(2):421–433, 2006.

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

[Faenza *et al.*, 2019] Yuri Faenza, Telikepalli Kavitha, Vladlena Powers, and Xingyu Zhang. Popular matchings and limits to tractability. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2790–2809, 2019.

[Fanelli *et al.*, 2021] Angelo Fanelli, Gianpiero Monaco, and Luca Moscardelli. Relaxed core stability in fractional hedonic games. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 182–188, 2021.

[Fioravanti *et al.*, 2023] Simone Fioravanti, Michele Flammini, Bojana Kodric, and Giovanna Varricchio. ε-fractional core stability in hedonic games. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[Gairing and Savani, 2019] Martin Gairing and Rahul Savani. Computing stable outcomes in symmetric additively separable hedonic games. *Mathematics of Operations Research*, 44(3):1101–1121, 2019.

[Gale and Shapley, 1962] David Gale and Lloyd S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

[Gärdenfors, 1975] Peter Gärdenfors. Match making: Assignments based on bilateral preferences. *Behavioral Science*, 20(3):166–173, 1975.

[Gupta *et al.*, 2021] Sushmita Gupta, Pranabendu Misra, Saket Saurabh, and Meirav Zehavi. Popular matching in roommates setting is np-hard. *ACM Transactions on Computation Theory*, 13(2):9:1–9:20, 2021.

[Kavitha *et al.*, 2011] Telikepalli Kavitha, J. Mestre, and M. Nasre. Popular mixed matchings. *Theoretical Computer Science*, 412(24):2679–2690, 2011.

[Kerkmann and Rothe, 2020] Anna M. Kerkmann and Jörg Rothe. Altruism in coalition formation games. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 461–467, 2020.

[Kerkmann *et al.*, 2020] Anna M. Kerkmann, Jérôme Lang, Anja Rey, Jörg Rothe, Hilmar Schadrack, and Lena Schend. Hedonic games with ordinal preferences and thresholds. *Journal of Artificial Intelligence Research*, 67:705–756, 2020.

[Newman, 2004] Mark E. J. Newman. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):321–330, 2004.

[Olsen, 2012] Martin Olsen. On defining and computing communities. In *Proceedings of the 18th Computing: The Australasian Theory Symposium (CATS)*, volume 128 of *Conferences in Research and Practice in Information Technology (CRPIT)*, pages 97–102, 2012.

[Papadimitriou, 1994] Christos H. Papadimitriou. *Computational Complexity*. Addison-Wesley, 1994.

[Peters, 2017] Dominik Peters. Precise complexity of the core in dichotomous and additive hedonic games. In *Proceedings of the 5th International Conference on Algorithmic Decision Theory (ADT)*, pages 214–227, 2017.

[Ray, 2007] Debraj Ray. *A Game-Theoretic Perspective on Coalition Formation*. Oxford University Press, 2007.

[Stockmeyer, 1977] Larry J. Stockmeyer. The polynomial-time hierarchy. *Theoretical Computer Science*, 3(1):1–22, 1977.

[Sung and Dimitrov, 2010] Shao C. Sung and Dinko Dimitrov. Computational complexity in additive hedonic games. *European Journal of Operational Research*, 203(3):635–639, 2010.

[Woeginger, 2013] Gerhard J. Woeginger. A hardness result for core stability in additive hedonic games. *Mathematical Social Sciences*, 65(2):101–104, 2013.