

# Non-Obvious Manipulability in Additively Separable and Fractional Hedonic Games

Diodato Ferraioli<sup>1</sup> and Giovanna Varricchio<sup>2</sup>

<sup>1</sup> University of Salerno, Italy

<sup>2</sup> University of Calabria, Italy

dferraioli@unisa.it, giovanna.varricchio@unical.it

## Abstract

In this work, we consider the design of Non-Obviously Manipulable (NOM) mechanisms, mechanisms that bounded rational agents may fail to recognize as manipulable, for two relevant classes of succinctly representable Hedonic Games: Additively Separable and Fractional Hedonic Games. In these classes, agents have cardinal scores towards other agents, and their preferences over coalitions are determined by aggregating such scores. This aggregation results in a utility function for each agent, which enables the evaluation of outcomes via the utilitarian social welfare. We first prove that, when scores can be arbitrary, every optimal mechanism is NOM; moreover, when scores are limited in a continuous interval, an optimal mechanism that is NOM exists. Given the hardness of computing optimal outcomes in these settings, we turn our attention to efficient and NOM mechanisms. To this aim, we first prove a characterization of NOM mechanisms that simplifies the class of mechanisms of interest. Then, we design a NOM mechanism returning approximations that asymptotically match the best-known approximation achievable in polynomial time. Finally, we focus on discrete scores, where the compatibility of NOM with optimality depends on the specific values. Therefore, we initiate a systematic analysis to identify which discrete values support this compatibility and which do not.

## 1 Introduction

*Hedonic Games* are a well-established model for describing coalition formation [Dreze and Greenberg, 1980]. In these games, agents have preferences over possible coalitions, and the goal is to compute a “good” partition of the agents, which constitutes the outcome of the game. Two approaches have been considered for the computation of these partitions: a decentralized one, in which agents autonomously choose their coalition; and a centralized one, where agents reveal their preferences to a designer who determines the outcome. While, in the first approach, it is assumed that the attained

partition satisfies some form of stability against either individual or group deviation [Bloch and Diamantoudi, 2011; Feldman *et al.*, 2015; Gairing and Savani, 2010; Bogomolnaia and Jackson, 2002; Banerjee *et al.*, 2001; Elkind and Wooldridge, 2009; Igarashi and Elkind, 2016], in the second, it is instead necessary to design algorithms, a.k.a., *mechanisms*, incentivizing agents to report their real preferences [Flammini *et al.*, 2021a; Flammini *et al.*, 2022; Klaus *et al.*, 2023; Dimitrov and Sung, 2004; Dimitrov *et al.*, 2006; Rodríguez-Álvarez, 2009; Varricchio, 2023].

In the design of these mechanisms, the goal is thus to align the scope of the designer, i.e., to compute partitions that satisfy some useful property, with the one of the agents, that is, being assigned to coalitions that they prefer. This alignment is usually achieved by requiring that the algorithm used for computing partitions satisfies a property named *strategyproofness* (SP). Strategyproofness ensures that an agent’s best outcome, given the preferences of other agents, is obtained by truthfully reporting her own preferences. Unfortunately, strategyproofness has been proved to be often a too demanding requirement: it indeed assumes that agents are fully rational, meaning that they would be able to evaluate which report would be the best for *every* possible realization of the game (i.e., for every possible set of preferences expressed by other agents). In some cases, this would be computationally unaffordable due to the exponentially large set of realizations to verify. In other cases, it requires an agent to parse very complex mathematical proofs, which is not feasible for agents with a scarce mathematical background. Moreover, it has been observed that in many settings agents lack contingent reasoning [Charness and Levin, 2009; Esponda and Vespa, 2014; Ngangoué and Weizsäcker, 2021], i.e., they are not inclined to execute case analysis, and they make their decision by aggregating different cases.

Moreover, asking for strategyproof mechanisms against fully rational agents usually causes a very bad quality of the computed partitions. In particular, Özyurt and Sanver [2009] show that it is in general impossible to have non-dictatorial strategyproof mechanisms. Similar impossibility results have also been provided in more restricted settings and specific coalitions’ quality measures [Amanatidis *et al.*, 2017; Brandl *et al.*, 2018]. For example, Flammini *et al.* [2021a] proved that no strategyproof mechanism can return a bounded approximation of the optimal partition, even for games whose

preferences are succinctly representable, and the partitions are evaluated with respect to the social welfare.

Hence, it would be very natural to ask whether better partitions can be achieved if one only requires that the mechanism is strategyproof against bounded rational agents or, more precisely, against agents lacking contingent reasoning skills. Recall that these agents are supposed to submit false preferences (i.e., to manipulate) only if this is convenient to them based on an aggregation of the different possible realizations. In particular, very recently, there has been a lot of attention on *Non-Obvious Manipulable* (NOM) mechanisms [Troyan and Morrill, 2020], that aggregate the realizations by only considering the best and worst cases. Specifically, in NOM mechanisms, an agent does not submit a false report only if: (i) the best outcome (over all possible reports of other agents) she can achieve for this report is not better than the best outcome that the agent can achieve by submitting her real preferences; and (ii) the worst outcome achievable with this report is not better than the worst outcome achievable with the true report. This choice for aggregating the possible realizations is motivated by empirical evidence from the failure of people in recognizing profitable manipulations, mainly in school choice and two-sided matching problems [Dur *et al.*, 2018; Pathak and Sönmez, 2008].

NOM mechanisms have been recently designed for different problems: school choice, two-sided matching, auctions, bilateral trade [Troyan and Morrill, 2020], voting [Aziz and Lam, 2021; Arribillaga and Bonifacio, 2024], fair division [Ortega and Segal-Halevi, 2022; Psomas and Verma, 2022], and single-parameter domains [Archbold *et al.*, 2023a; Archbold *et al.*, 2023b; Archbold *et al.*, 2024]. For many of these, it has been shown that NOM mechanisms can produce much better outcomes than SP mechanisms. Very recently, similar results have been obtained by Flammini *et al.* [2025] for hedonic games with friends’ appreciation preferences, a special class of additively separable hedonic games. Hence, the hope is that the same may occur also for more expressive classes of hedonic games.

**Our Contribution.** We focus on two well-studied subclasses of hedonic games with succinct preference representation: *Fractional Hedonic Games* (FHG) [Aziz *et al.*, 2019] and *Additively Separable Hedonic Games* (ASHG) [Hajduková, 2006]. Here, preferences over coalitions are expressed by utilities computed from the scores assigned by an agent to every other coalition member. This allows evaluating the partition’s quality by using *utilitarian social welfare*, the sum of agents’ utilities for the outcome.

We first show that, while no SP mechanism may return even a bounded approximation of the optimal social welfare [Flammini *et al.*, 2021a], every mechanism that returns such the optimum is NOM. Unfortunately, this result does not settle the problem at all for two reasons. First, in both FHGs and ASHG, computing the social welfare maximizing partition is a computationally hard problem [Flammini *et al.*, 2022; Aziz *et al.*, 2015], and hence one would be interested in evaluating the approximation of the social welfare achievable by mechanisms that are both NOM and computationally efficient. Moreover, optimal mechanisms may fail to be NOM

as soon as scores are restricted to limited intervals.

We address the first issue by proving that an efficient NOM mechanism that returns an  $n$ -approximation of the optimal social welfare in ASHG, where  $n$  is the number of agents, and a 2-approximation in FHG exists. We stress that for ASHG the bound is tight, as an approximation better than  $n$  cannot be computed in poly-time, unless  $P=NP$ , even if one does not care about manipulability [Flammini *et al.*, 2022].

On the route to achieve this, we prove another result that may be of independent interest: We show that when one cares about NOM mechanisms with bounded approximation, it is without loss of generality to look at “natural” mechanisms, i.e., mechanisms that determine whether two agents must belong to the same coalition, only by considering how much they like/dislike each other (in total), regardless of the specific mutual scores. This characterization also allows us to show that, while not every optimal mechanism is NOM, there always exists an optimal mechanism that is NOM, even if scores are bounded in a continuous interval.

In turn, such a result no longer holds true as soon as we turn the attention to discrete values: for example, in [Flammini *et al.*, 2025] it has been proven that in ASHG, if scores are in  $\{-n, 1\}$ , then no optimal mechanism is NOM. In this work, we elaborate more on these settings by providing an almost full characterization of the existence of optimal NOM mechanisms for ASHG when scores can take values in  $\{-x, 0, 1\}$ , with  $x \geq 1$ . Interestingly, we show that, when  $x > 2n-3$ , every optimal mechanism is NOM, whereas no optimal mechanism is NOM for  $x \in (1, 2n-3)$ . However, for  $x = 1, 2n-3$ , NOM is possibly compatible with optimality, delineating that there is not a clear phase transition over the values of  $x$ .

Due to space limits, we defer some proofs to the full paper.

## 2 Notation and Preliminaries

This section presents the notation and preliminary results.

### 2.1 Hedonic Games

In *Hedonic Games* (HG) we are given a set of  $n$  agents, denoted by  $\mathcal{N} = \{1, \dots, n\}$ , and the goal is to split them into disjoint coalitions, that is, to provide a *partition*  $\pi = \{C_1, \dots, C_k\}$  such that  $\cup_{h=1}^k C_h = \mathcal{N}$ . Such a partition  $\pi$  is also called *outcome* or *coalition structure*. We might denote by  $\pi(i)$  the coalition  $i \in \mathcal{N}$  is assigned to in a partition  $\pi$ . The *grand coalition* refers to a partition that consists of one coalition containing all the agents; a *singleton coalition* refers to a coalition containing only one agent. The main characteristic of HGs is that agents have preferences over the possible outcomes that solely depend on the coalition they are assigned to and not on how the others aggregate. We denote the set of all possible coalitions of  $i$  by  $\mathcal{N}_i = \{C \subseteq \mathcal{N} | i \in C\}$ .

Of particular interest are HGs classes where agents express their preferences using utilities and admit succinct representation, meaning the memory used to store the necessary information to compute agents’ preferences is polynomial in the instance size. In fractional (FHGs) and additively separable (ASHGs) HGs, agents report individual scores towards other agents, and the utility they derive from the coalition they are assigned to is based on the value they attribute to the other

coalition members. More formally, each agent  $i$  has a weight function  $w_i : \mathcal{N} \rightarrow \mathbb{R}$  that determines the score that agent  $i$  assigns to any other participant. We always assume that  $w_i(i) = 0$  and denote  $w_{ij} = w_i(j)$ . In ASHG and FHGs, the *utility* that agent  $i$  derives in being in a coalition  $C \in \mathcal{N}_i$ , is given by

$$u_i(C) = \sum_{j \in C} w_{ij} \quad \text{and} \quad u_i(C) = \frac{\sum_{j \in C} w_{ij}}{|C|}, \text{ respectively.}$$

Note that  $u_i(C)$  is well-defined only if  $C \in \mathcal{N}_i$ .

In general,  $w_{ij}$  can be any value in  $\mathbb{R}$ . Nonetheless, we may also consider restrictions to specific settings. Specifically, we say that the agents' weights towards the other agents are:

- *arbitrary* if  $w_{ij} \in \mathbb{R}$ , for all  $i, j \in \mathcal{N}$ ;
- *non-negative* if  $w_{ij} \in \mathbb{R}_{\geq 0}$ , for all  $i, j \in \mathcal{N}$ ;
- *bounded* if  $w_{ij} \in [-1, 1]$ , for all  $i, j \in \mathcal{N}$ ;
- *general duplex* if  $w_{ij} \in \{-x, 0, 1\}$ , for all  $i, j \in \mathcal{N}$  and  $x \in (0, \infty)$  (in case  $x = 1$  we simply say *duplex*).

The distinction in arbitrary, non-negative, bounded, and duplex weights has been presented in [Flammini *et al.*, 2021a]. We introduce the notion of general duplex valuations to include the well-known class of friends and enemies games [Dimitrov and Sung, 2004]. If we do not specify otherwise, we assume the values are arbitrary real numbers.

**Graph Representation and Flattened Graph.** ASHG and FHGs have a very convenient representation by means of directed and weighted graphs  $\vec{G} = (\mathcal{N}, \{w_i\}_{i \in \mathcal{N}})$ , where  $\mathcal{N}$  are the vertices and a directed edge  $(i, j)$  has weight  $w_{ij}$ , denoting the value  $i$  has for agent  $j$ . We always assume the graphs to be complete: an edge exists even if its weight is 0.

Given two disjoint coalitions  $C_1, C_2 \subseteq \mathcal{N}$ , we refer to the *cut*  $(C_1, C_2)$  induced in the graph  $\vec{G}$  by these two coalitions, that is,  $Cut(C_1, C_2) = \{(i, j), (j, i) \mid i \in C_1, j \in C_2\}$ . The value of the cut  $Cut(C_1, C_2)$  is  $\sum_{i \in C_1, j \in C_2} (w_{ij} + w_{ji})$ . Moreover, we denote by  $\delta^+(i) = \{j \neq i \mid w_{ij} > 0\}$  and  $\delta^-(i) = \{j \neq i \mid w_{ij} < 0\}$  the set of positive and negative neighbors of  $i$ , respectively.

When focusing on the utilitarian social welfare of a partition, given a pair of agents  $i, j$ , the values  $w_{ij}$  and  $w_{ji}$  either both contribute to the social welfare or neither does. Therefore, we may consider an undirected version of  $\vec{G}$  where the edge weight of an edge is given by the sum of the mutual values. More formally, we say that a weighted and undirected graph  $G = (\mathcal{N}, \hat{w})$  is the *flattened graph* of a directed and weighted graph  $\vec{G}$ , if it has the same vertices of  $\vec{G}$  and, for each pair of directed edges  $(i, j)$  and  $(j, i)$  of weights  $w_{ij}$  and  $w_{ji}$  in  $\vec{G}$ , respectively, there is an undirected edge  $e = \{i, j\}$  of weight  $\hat{w}(e) = \hat{w}(i, j) = w_{ij} + w_{ji}$  in  $G$ . Notice that an ASHG or FHG instance is uniquely determined by the directed graph  $\vec{G}$  we described above; however, different instances may have the same flattened graph  $G$ .

**Proportional Graphs.** Given two undirected and weighted graphs  $G$  and  $G'$  with the same set of vertices, and having edge weights  $\hat{w}$  and  $\hat{w}'$ , we say that  $G$  and  $G'$  are *proportional* if there exists  $\lambda > 0$  such that  $\hat{w}(i, j) = \lambda \cdot \hat{w}'(i, j)$

holds true for each pair of vertices  $i \neq j$ . Notice that proportionality is an equivalence relation over the set of weighted and undirected graphs; we write  $G' \sim_P G''$  to denote that the two graphs are proportional. We often slightly abuse notation, and we will simply say that two instances, namely  $\mathcal{I}, \mathcal{I}'$ , are proportional, and write  $\mathcal{I} \sim_P \mathcal{I}'$ , if the flattened graphs modeling these instances are proportional. In Section 4.2, proportionality will turn out to be a useful property to characterize mechanisms resistant to obvious manipulations.

## 2.2 Strategyproofness and Manipulability

In order to compute desirable partitions of the agents, we need to know their preferences. However, the values  $w_{ij}$  might be private information of agent  $i$ , so agents must reveal them. Let  $\mathbf{d} = (d_1, \dots, d_n)$  be the vector of agents' declarations, where the declaration  $d_i$  contains the information revealed by agent  $i$ : by assuming direct revelation,  $d_i = \{d_{ij}\}_{j \in \mathcal{N} \setminus \{i\}}$ , where  $d_{ij}$  represents the value that agent  $i$  declares for agent  $j$ . We denote by  $\mathcal{D}$  the space of all possible declarations. For convenience,  $\mathbf{d}_{-i}$  denotes the declarations of all agents except  $i$ , and  $\mathcal{D}_{-i}$  is the set of all feasible  $\mathbf{d}_{-i}$ . Similarly, we denote by  $\mathcal{D}_i$  the space of feasible  $d_i$ .

We shall denote by  $\mathcal{M}$  a mechanism and by  $\mathcal{M}(\mathbf{d})$  its output upon a declaration  $\mathbf{d} \in \mathcal{D}$  of the agents. We denote by  $\mathcal{M}_i(\mathbf{d})$  the coalition  $i$  is assigned to. Being our agents strategic, an agent  $i$  could possibly declare  $d_i \neq w_i$ , where  $w_i$  contains the real values assigned by agent  $i$ , and will be henceforth also called *real type*. In this context, a key challenge is to design algorithms – commonly referred to as *mechanisms* – encouraging agents to behave truthfully, i.e., to submit a declaration  $d_i = w_i$  where she reveals her true values towards the other agents. The most desirable and widely studied characteristic for such a kind of mechanisms is *strategyproofness*.

**Definition 1** (Strategyproofness). A mechanism  $\mathcal{M}$  is said to be *strategyproof* (SP) if for each  $i \in \mathcal{N}$ , and any declaration of the other agents  $\mathbf{d}_{-i}$

$$u_i(\mathcal{M}(w_i, \mathbf{d}_{-i})) \geq u_i(\mathcal{M}(d_i, \mathbf{d}_{-i})) \quad (1)$$

holds true for any possible declaration  $d_i \neq w_i$  of agent  $i$ .

In turn, a mechanism is said to be *manipulable* if there exists an agent  $i$ , a real type  $w_i$ , and a declaration  $d_i \neq w_i$  s.t. Eq. 1 does not hold true. Such a  $d_i$  is called a *manipulation*.

Due to its demanding requirement, strategyproof mechanisms have been shown to be highly inefficient; for this reason, mechanisms satisfying milder conditions have been introduced with the scope of obtaining more efficient outcomes.

**Definition 2** (Non-Obvious Manipulability). A mechanism  $\mathcal{M}$  is said to be *non-obviously manipulable* (NOM) if for every  $i \in \mathcal{N}$ , real type  $w_i$ , and any other declaration  $d_i$  the following two conditions hold true:

1.  $\sup_{\mathbf{d}_{-i}} u_i(\mathcal{M}(w_i, \mathbf{d}_{-i})) \geq \sup_{\mathbf{d}_{-i}} u_i(\mathcal{M}(d_i, \mathbf{d}_{-i}))$ , and
2.  $\inf_{\mathbf{d}_{-i}} u_i(\mathcal{M}(w_i, \mathbf{d}_{-i})) \geq \inf_{\mathbf{d}_{-i}} u_i(\mathcal{M}(d_i, \mathbf{d}_{-i}))$ .

In case there exist  $i, w_i$ , and  $d_i$  such that one of the aforementioned conditions is violated, then  $\mathcal{M}$  is *obviously manipulable* and  $d_i$  is an *obvious manipulation*.

In other words, to determine if a mechanism  $\mathcal{M}$  is NOM for an agent  $i$ , we need to consider the best- and the worst-case scenario for  $i$  (according to her truthful preferences) of  $\mathcal{M}$  when  $i$  declares her true type  $w_i$  and any other possible declaration  $d_i \neq w_i$ . In particular, neither the best nor the worst possible outcome when declaring  $w_i$  can be strictly worse than the best or the worst outcome, according to her true preferences, attainable when declaring  $d_i$ .

Clearly, for any game class,  $\text{SP} \Rightarrow \text{NOM}$ . That is, SP is more demanding than NOM as it requires verifying for every declaration of the others if there exists a profitable manipulation. However, the agents might not know upfront what the others' declarations are. This is circumvented by NOM.

In order to work with non-obvious manipulability, it is necessary to understand the structure of outcomes guaranteeing such a property. We here provide a first result in this direction: a simple but still useful sufficient condition for a mechanism being NOM. A finer characterization will be given later.

To this aim, let us introduce some necessary notation. Given a mechanism  $\mathcal{M}$ , for any  $d_i \in \mathcal{D}_i$ , let

$$\text{Out}_i^{\mathcal{M}}(d_i) = \{\mathcal{M}(d_i, \mathbf{d}_{-i}) \mid \mathbf{d}_{-i} \in \mathcal{D}_{-i}\}$$

be the set of possible outcomes if  $i$  declares  $d_i$ . Moreover,

$$\text{Coal}_i^{\mathcal{M}}(d_i) = \{\mathcal{M}_i(d_i, \mathbf{d}_{-i}) \mid \mathbf{d}_{-i} \in \mathcal{D}_{-i}\}$$

is the set of possible coalitions for  $i$  if she declares  $d_i$ .

**Proposition 1.** *Given a mechanism  $\mathcal{M}$ , if for each  $i \in \mathcal{N}$  and  $d_i \in \mathcal{D}_i$ , given the truthful declaration  $w_i$ ,*

1.  $\text{Coal}_i^{\mathcal{M}}(w_i) = \text{Coal}_i^{\mathcal{M}}(d_i)$  or
2.  $\text{Out}_i^{\mathcal{M}}(w_i) = \text{Out}_i^{\mathcal{M}}(d_i)$ ,

then,  $\mathcal{M}$  is NOM.

*Proof.* Suppose first that  $\text{Coal}_i^{\mathcal{M}}(w_i) = \text{Coal}_i^{\mathcal{M}}(d_i)$ . That is, whatever the declaration of agent  $i$  is, the set of coalitions  $i$  may end up with remains the same. Then, the best/worst outcome remains the same as well, and the conditions (1) and (2) in Definition 2 are always satisfied with the equality, and thus NOM holds true. Note that  $\text{Out}_i^{\mathcal{M}}(w_i) = \text{Out}_i^{\mathcal{M}}(d_i)$  implies  $\text{Coal}_i^{\mathcal{M}}(w_i) = \text{Coal}_i^{\mathcal{M}}(d_i)$ , and thus the thesis follows even for the condition 2.  $\square$

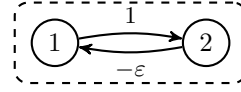
Proposition 1 provides two simple yet useful sufficient conditions for determining whether a mechanism is NOM.

### 2.3 Optimal and Approximate Solutions

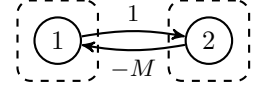
Being the agents' preferences expressed by means of utilities, we can evaluate the quality of outcomes through a *social welfare* (SW). Specifically, let  $\pi$  be any outcome; the (utilitarian) SW of  $\pi$  is given by  $\text{SW}(\pi) = \sum_{i \in \mathcal{N}} u_i(\pi(i))$ . We call *social optimum*, or simply the optimum, any outcome  $\text{OPT}$  in  $\arg \max_{\pi} \text{SW}(\pi)$  and we denote by  $\text{opt}$  the value  $\text{SW}(\text{OPT})$ .

In the considered classes of games, computing an optimal outcome is often intractable. Therefore, we may focus on finding approximate solutions.

Let  $\text{opt}_{\mathbf{d}}$  the value of the social optimum for the declaration  $\mathbf{d}$ , i.e.,  $\text{opt}_{\mathbf{d}} = \max_{\pi} \sum_{i \in \mathcal{N}} \sum_{j \in \pi(i)} d_{ij}$  for ASHG, and  $\text{opt}_{\mathbf{d}} = \max_{\pi} \sum_{i \in \mathcal{N}} \frac{\sum_{j \in \pi(i)} d_{ij}}{\pi(i)}$  for FHG.



(a) Truthful declaration.



(b) Manipulation of Agent 2.

Figure 1: No BAPX is SP. Dashed lines represent the only BAPX partition for the given instance.

**Definition 3** (Bounded approximation property (BAPX)). A mechanism  $\mathcal{M}$  satisfies the *bounded approximation property* (BAPX) if

$$1 \leq \sup_{\mathbf{d} \in \mathcal{D}} \frac{\text{opt}_{\mathbf{d}}}{\text{SW}(\mathcal{M}(\mathbf{d}))} \leq r(n)$$

holds true for a real-valued and bounded function  $r(n)$ . We denote by  $r^{\mathcal{M}}$  the approximation ratio of a mechanism  $\mathcal{M}$ .

If  $\text{opt}_{\mathbf{d}} = 0$  the only bounded approximation mechanism is the one returning the optimum.<sup>1</sup>

We ask our mechanisms to return solutions that guarantee a non-negative SW, since there is at least one non-negative solution (namely, each agent in a singleton), and hence also the optimum is non-negative.

A mechanism that for any input instance returns an optimum partition is called *optimal*.

As we already mentioned, strategyproofness turned out to be extremely inefficient in terms of approximation to the social optimum. For the sake of completeness, we include the next theorem, which is a restatement of results from [Flammini et al., 2021a].

**Proposition 2.** *For both ASHGs and FHGs with arbitrary weights, no SP mechanism is BAPX.*

*Sketch.* The following applies to both ASHGs and FHGs.

Assume there are two agents and the graph depicted in Figure 1a describes their truthful preferences. Any BAPX mechanism must put the agents in the same coalition. However, this outcome guarantees to agent 2 a negative utility.

In turn, if agent 2 misreports, as shown in Figure 1b, the only way to achieve BAPX in this instance is to put the agents in singletons. This guarantees to agent 2 a higher utility and shows there exists a manipulation.  $\square$

Fortunately, NOM stands in contrast to this impossibility result. We will demonstrate that NOM is, in fact, compatible with optimality, and that any approximation algorithm with BAPX can be transformed into a NOM mechanism that preserves the same approximation guarantee.

### 3 Optimality and NOM

While strategyproofness is incompatible with *any* bounded approximation of the optimum (see Proposition 2), in this section, we show that non-obvious manipulability is instead compatible with optimality, at least for arbitrary weights.

**Theorem 1.** *For both ASHGs and FHGs with arbitrary weights, any optimal mechanism is NOM.*

<sup>1</sup>In this case, we assume “ $\frac{0}{0} = 1$ ”.

*Proof.* The next arguments hold for both ASHG and FHGs.

Let us fix an agent  $i$ , and let  $C \in \mathcal{N}_i$ . We next show that for any  $d_i \in \mathcal{D}_i$  there exists  $\mathbf{d}_{-i} \in \mathcal{D}_{-i}$  such that, in any optimal partition  $\pi$  of instance  $\mathbf{d} = (d_i, \mathbf{d}_{-i})$ , it holds  $\pi(i) = C$ .

Since the weights range in  $\mathbb{R}$ , for any  $d_i$  we can select  $\mathbf{d}_{-i} \in \mathcal{D}_{-i}$  in such a way that the flattened graph of the instance,  $G = (\mathcal{N}, \hat{w})$ , has the weights defined as follows:  $\hat{w}(e) = 1$ , for all  $e = (j, j')$  with  $j, j' \in C$ , and  $\hat{w}(e) = -M$ , for a sufficiently large  $M$ , otherwise.

Let  $\pi$  be any optimal partition for such an instance, and suppose that there is  $j \in \pi(i)$  such that  $j \notin C$ ; then,  $\text{SW}(\pi(i) \setminus \{j\}) - \text{SW}(\pi(i)) < 0$ , for an  $M$  large enough. Being such a difference negative, it would be preferable to move  $j$  to a singleton coalition, contradicting the optimality of  $\pi$ . Therefore,  $\pi(i) \subseteq C$ .

Assume there exist two disjoint coalitions in  $\pi$ ,  $C', C''$ , both contained in  $C$ . Being  $\pi$  optimum,  $\Delta = \text{SW}(C' \cup C'') - \text{SW}(C') - \text{SW}(C'') \leq 0$ , as it must not be strictly more convenient to merge these two coalitions for the SW. However, in the case of ASHG,  $\Delta$  measures the total weight of the cut  $(C', C'')$ . Being this cut made of positive edges only, we have  $\Delta > 0$ , a contradiction to the optimality of  $\pi$ . In FHGs,  $\Delta = (|C' \cup C''| - 1) - (|C'| - 1) - (|C''| - 1) > 0$ , again a contradiction to the optimality of  $\pi$ .

In conclusion, for any optimal  $\mathcal{M}$ ,  $\text{Coal}_i^{\mathcal{M}}(d_i) = \mathcal{N}_i$ , for each  $d_i \in \mathcal{D}_i$ , including the truthful declaration  $w_i$  of  $i$ . Therefore, the condition 1 of Proposition 1 is satisfied by any optimal mechanism, and our claim follows.  $\square$

Theorem 1 poses NOM in sharp contrast with SP. Unfortunately, such a result does not hold for any class of weights as the following example shows.

**Example 1.** Assume weights are bounded. The following arguments apply to both ASHG and FHGs.

Let  $\mathcal{M}$  be an optimal mechanism that, whenever there are two agents and  $d_{12} + d_{21} = 0$ , works as follows: if  $d_{12} = -1$ , the mechanism puts the two agents in distinct coalitions, otherwise, the two agents are put in the same coalition. Note that if  $d_{12} + d_{21} = 0$ , any partition of the two agents is optimal.

If the true type  $w_1$  of agent 1 is such that the score towards agent 2 is  $-x$ , with  $x < 1$ , then,  $\text{Out}_1^{\mathcal{M}}(w_1) = \{\{\mathcal{N}\}, \{\{1\}, \{2\}\}\}$ . In fact, whenever  $d_{21} > x$  (which is possible as long as  $x < 1$ ) the only optimum is achieved by putting the agents in the same coalition; in turn, if  $d_{21} < x$  the optimum must split the agents in separate coalition.

Assume agent 1 misreports by declaring  $d_1$  where  $d_{12} = -1$ . In this case,  $\text{Out}_1^{\mathcal{M}}(d_1) = \{\{\{1\}, \{2\}\}\}$ . In fact,  $d_{12} + d_{21} \leq 0$  as  $d_{21} \leq 1$ , and hence, if  $d_{12} + d_{21} < 0$  any optimal outcome will put the agents in singletons, otherwise, the mechanism will apply the tie-breaking rule described above, showing that the only possible partition is  $\{\{1\}, \{2\}\}$ . Therefore, this manipulation improves the worst case for 1, proving that the optimal mechanism  $\mathcal{M}$  is not NOM.

Even though optimal mechanisms may not necessarily be NOM, in the next section, we provide a characterization of BAPX and NOM mechanisms. As a result, this will show that there always exists an optimal and NOM mechanism, both for bounded and non-negative weights.

## 4 BAPX and NOM Mechanisms

In the previous section, we observed that while all optimal mechanisms are NOM, when weights range in  $\mathbb{R}$ , this is not necessarily the case when weights are bounded. However, looking back to Example 1, we highlight that some strange behavior of the mechanism caused its obvious manipulability. In particular, such a mechanism may return different optimal outcomes for two proportional instances (instances having the same flattened graph).

We notice that, given any outcome  $\pi$ , whenever two agents  $i, j$  are in the same coalition, both  $w_{ij}$  and  $w_{ji}$  contribute to the social welfare; conversely, if they are not in the same coalition, neither  $w_{ij}$  nor  $w_{ji}$  contributes to it. For this reason, the optimality of an outcome is *scale-independent*, that is, if an outcome is optimum, it remains optimum for a proportional instance. Leading to the following observation.

**Observation 1.** In ASHG and FHGs, proportional instances have the same optimal outcomes.

Therefore, since by Observation 1 proportional instances share the same optimal solutions, it looks reasonable that, when designing an optimal mechanism, one should care only about the flattened graph of the instance, rather than the mutual evaluations of the agents. For all these reasons, we introduce the property of *scale-independence* for a mechanism.

**Definition 4** (Scale-Independence (SI)). A mechanism  $\mathcal{M}$  is *scale-independent* (SI) if for any pair of proportional inputs,  $\mathbf{d}, \mathbf{d}'$ , it returns the same outcome, that is,  $\mathcal{M}(\mathbf{d}) = \mathcal{M}(\mathbf{d}')$ .

SI requires the mechanism to be consistent with the flattened graph, and any graph proportional to it, meaning that the outcome should solely depend on the contribution of pairs of agents to the social welfare, rather than the individual contributions of the agents. This assumption looks quite natural, as the mutual values of a pair of agents contribute to the social welfare only if the agents are in the same coalition.

### 4.1 A Sufficient Condition for NOM

In Example 1, the presented optimal mechanism is not NOM and does not even satisfy SI. With the next theorem, we further emphasize the importance of considering SI mechanisms as this suffices to guarantee NOM.

**Theorem 2.** For both ASHG and FHGs with arbitrary, non-negative, or bounded weights,  $\text{SI} \Rightarrow \text{NOM}$ .

*Sketch.* The bulk of our proof is given by the following observation, the proof can be found in the full paper. If  $\mathcal{D}$  is the space of arbitrary, bounded, or non-negative weights, then, for each  $i \in \mathcal{N}$ , the following holds true:

- ( $\star$ ) for any  $d_i, d'_i \in \mathcal{D}_i$ , and  $\mathbf{d}_{-i} \in \mathcal{D}_{-i}$ , there exists  $\mathbf{d}'_{-i} \in \mathcal{D}_{-i}$  such that  $(d_i, \mathbf{d}_{-i}) \sim_P (d'_i, \mathbf{d}'_{-i})$ .

Let  $\mathcal{M}$  be an SI mechanism, and  $\pi = \mathcal{M}(d_i, \mathbf{d}_{-i})$  be an outcome of  $\text{Out}_i^{\mathcal{M}}(d_i)$ . Being  $\mathcal{M}$  scale-independent, and given that ( $\star$ ) holds true, then, for any other  $d'_i \in \mathcal{D}_i$  there exists  $\mathbf{d}'_{-i} \in \mathcal{D}_{-i}$  such that  $\mathcal{M}(d_i, \mathbf{d}_{-i}) = \mathcal{M}(d'_i, \mathbf{d}'_{-i})$ , and thus  $\pi \in \text{Out}_i^{\mathcal{M}}(d'_i)$ . This shows that for any pair of  $d_i, d'_i$ ,  $\text{Out}_i^{\mathcal{M}}(d_i) = \text{Out}_i^{\mathcal{M}}(d'_i)$ . Therefore, by choosing  $d'_i = w_i$ , Proposition 1 proves  $\mathcal{M}$  is NOM.  $\square$

## 4.2 A Characterization of BAPX and NOM

Let us now specifically focus on BAPX mechanisms. Notice that since optimal mechanisms may not be SI (see Example 1), the same holds true for BAPX mechanisms.

We next show that, for any possible approximation ratio  $\rho$ , there always exists an SI mechanism with approximation factor exactly  $\rho$ . To this end, we will exploit some interesting properties deriving from the definition of the equivalence class  $\sim_P$  over game instances. In particular, we introduce a poly-time procedure which, for any pair of instances  $\mathcal{I}, \mathcal{I}'$  such that  $\mathcal{I} \sim_P \mathcal{I}'$ , selects the same representative for their equivalence class. We call this procedure REPR and it works as follows: Given a game instance  $\mathcal{I}$ , it builds the corresponding flattened graph  $G \in \mathcal{G}^n$ ; next, REPR transform  $G = (\mathcal{N}, \hat{w})$  into a proportional flattened graph  $G' = (\mathcal{N}, \hat{w}')$  such that the maximum weight in absolute terms in  $G'$  equals 1. This is clearly always possible by normalizing weights by a factor  $\max_{i \neq j} |\hat{w}(i, j)|$ ; finally, by assuming that there is a prefixed ordering of the agents, say  $1, \dots, n$ , REPR returns an instance with agents  $\mathcal{N}$  and weights  $w''_i$ , for all  $i \in \mathcal{N}$ , defined as follows:

$$w''_{ij} = \begin{cases} \hat{w}'(i, j) & \text{if } i < j \\ 0 & \text{otherwise.} \end{cases}$$

We will call the instance returned by  $\text{REPR}(\mathcal{I})$  as the *representative instance* of  $\mathcal{I}$ .

It is not hard to see that  $\text{REPR}(\mathcal{I}) \sim_P \mathcal{I}$ , and for any pair of proportional instances  $\mathcal{I}, \mathcal{I}'$  we have  $\text{REPR}(\mathcal{I}) = \text{REPR}(\mathcal{I}')$ . Moreover, the procedure REPR works in polynomial time.

REPR turns out to be very helpful in transforming a mechanism into an SI mechanism while maintaining the same approximation guarantee, as the following lemmas show.

**Lemma 1.** *For any mechanism  $\mathcal{M}$ , the mechanism  $\mathcal{M}' = \mathcal{M} \circ \text{REPR}^2$  is SI.*

**Lemma 2.** *If  $\mathcal{M}$  is a  $\rho$ -approximating mechanism, then,  $\mathcal{M}' = \mathcal{M} \circ \text{REPR}$  is a  $\rho$ -approximating mechanism as well.*

Lemmas 1 and 2 together with Theorem 2 show the following theorem.

**Theorem 3.** *For any  $\rho \geq 1$ , in ASHG or FHGs with arbitrary, non-negative, or bounded weights, there exists a NOM and  $\rho$ -approximating mechanism if and only if there exists an SI and  $\rho$ -approximating mechanism.*

Essentially, Theorem 3 allows us to focus our attention on generic  $\rho$ -approximating mechanisms. Indeed, each such algorithm can be turned into a  $\rho$ -approximating and SI mechanism, through the procedure REPR, and hence in a  $\rho$ -approximating and NOM mechanism, by Theorem 2.

## 4.3 Approximate NOM Mechanism

We recall that, in ASHG, whenever the agents' weights are non-negative, the optimum, which consists in grouping all the agents in the same coalition, is SP. In turn, no SP mechanism has a bounded approximation ratio for arbitrary and bounded weights. Moreover, regardless of the SP or NOM

requirement, no algorithm can provide an approximation in ASHG that is  $O(n^{1-\varepsilon})$ , for every  $\varepsilon > 0$ , even in simple cases [Flammini *et al.*, 2022], while computing welfare maximizing partitions in FHGs is NP-hard [Aziz *et al.*, 2015]. In what follows, we focus on determining NOM and efficient mechanisms for both arbitrary, non-negative, and bounded weights.

**Mechanism  $\mathcal{M}_1$ .** Given an instance  $\mathcal{I}$ , let  $\mathcal{M}$  be a mechanism that builds the flattened graph corresponding to  $\mathcal{I}$ , computes a maximum weighted matching on it, and creates a coalition for each matched pair of agents; unmatched agents are put in singletons. Then, we define  $\mathcal{M}_1 = \mathcal{M} \circ \text{REPR}$ .

**Theorem 4.**  *$\mathcal{M}_1$  is NOM for both ASHG and FHGs under unbounded, non-negative, and bounded weights. Furthermore, it provides an  $n$ -approximation of  $\text{opt}$  for ASHG and a 2-approximation for FHGs.*

To show the theorem, we will make use of the following well-known fact (see e.g. Problem 16.5 of [Soifer, 2009]):

**Fact 1.** Let  $K$  be a clique of size  $k$ . Its edges can be partitioned into  $k-1$ , if  $k$  is even, and  $k$ , if odd, disjoint matchings.

*Proof of Theorem 4.* (NOM) By Lemma 1,  $\mathcal{M}_1$  is SI; therefore, from Theorem 2,  $\mathcal{M}_1$  is NOM for instances with arbitrary, non-negative, or bounded weights.

(Approximation) Let  $\mathcal{M}$  be the subroutine as described in the definition of  $\mathcal{M}_1$ . We next show the approximation guarantee of  $\mathcal{M}$  for both ASHG and FHGs. Being  $\mathcal{M}_1 = \mathcal{M} \circ \text{REPR}$ , by Lemma 2, this approximation guarantee is preserved in  $\mathcal{M}_1$ . Let  $\pi$  denote the outcome of  $\mathcal{M}$ .

Consider an optimal partition OPT and let  $k$  be the number of its coalitions, say,  $C_1, \dots, C_k$ . Let us compute for each coalition  $C_h$  a maximum matching  $M_h$  whose edges are in the subgraph by considering only the vertices in  $C_h$ , and let  $v_h$  be the value of this matching (the sum of its edges weights). Let us now consider a coalition in the optimum  $C_h$  of size  $|C_h| = c_h$ . By Fact 1, the edges of  $C_h$  can be partitioned into either  $c_h$  or  $c_h - 1$  disjoint matchings. Each of these matchings has a value of at most  $v_h$ , due to the maximality of  $M_h$ . Therefore,  $\sum_{i,j \in C_h} w_{ij} \leq c_h \cdot v_h \leq n \cdot v_h$ .

Consider now the maximum matching  $M^*$  computed by  $\mathcal{M}$ , and let us denote by  $v^*$  its total value (the sum of the weights of its undirected edges).

In the case of ASHG,  $\text{SW}(C_h) = \sum_{i,j \in C_h} w_{ij}$  and being  $\text{opt} = \sum_{h=1}^k \text{SW}(C_h) \leq \sum_{h=1}^k n \cdot v_h \leq n \cdot v^* = n \cdot \text{SW}(\pi)$ , the approximation follows. Notice that the last inequality holds true as  $M = \cup_{h=1}^k M_h$  is a matching in  $G$ , and being  $M^*$  optimal, the value of  $M$  cannot be larger than  $v^*$ .

In FHGs,  $\text{SW}(C_h) = \frac{\sum_{i,j \in C_h} w_{ij}}{C_h} \leq v_h$ , and hence  $\text{opt} = \sum_{h=1}^k \text{SW}(C_h) \leq \sum_{h=1}^k v_h \leq v^* = 2 \cdot \text{SW}(\pi)$ . Notice that if the value of a matching is  $v$ , then, the SW of the corresponding partition equals  $v/2$ .  $\square$

## 5 NOM and Optimality in Discrete Settings

So far we have seen that when weights range in  $\mathbb{R}, \mathbb{R}_{\geq 0}$ , or  $[-1, 1]$ , optimality and NOM are compatible. However, in [Flammini *et al.*, 2025] it has been shown that for ASHG with weights in  $\{-n, 1\}$ , no optimal mechanism is NOM,

<sup>2</sup>We denote  $f(g(\cdot))$  by  $f \circ g(\cdot)$ .

while, if the weights are in  $\{-\frac{1}{n}, 1\}$  there exists a NOM and optimal mechanism. Therefore, even in simple discrete settings, the compatibility of NOM with optimality depends on the specific values the weights may take.

In what follows, we will dive into these known results to provide a more complete understanding of the compatibility of NOM with optimality in discrete settings. Since known results only focus on ASHG, we here only consider this class, and we assume general duplex valuations, i.e., weights that may take values in  $\{-x, 0, 1\}$  for  $x \in [1, \infty)$ , that contains and extends the setting considered in [Flammini *et al.*, 2025].

We are able to prove the following characterization.

**Theorem 5.** *For general duplex valuations with  $x \geq 1$ , the following hold true:*

- For  $x > 2n - 3$ , any optimal mechanism is NOM;
- for  $x \in (1, 2n - 3)$ , no optimal mechanism is NOM;
- for  $x \in \{1, 2n - 3\}$  there exists an optimal and NOM mechanism.

Note that for  $x \in (0, 1)$  the compatibility of NOM with optimality remains open, besides the case  $x = \frac{1}{n}$  [Flammini *et al.*, 2025]. We believe this case is worth further investigation.

We next sketch the proof of some interesting cases of Theorem 5. Hereafter, we denote by  $\Delta_i^+ = \delta_i^+(w_i) \cup \{i\}$  and  $\Delta_i^- = \delta_i^-(w_i) \cup \{i\}$ , where  $w_i$  is the truthful type of  $i$ .

**Case  $x > 2n - 3$ : any optimal mechanism is NOM.** To this end, we will make use of the following lemma.

**Lemma 3.** *Under general duplex valuations with  $x > 2n - 3$ , for any optimal partition  $\pi$ , there is no negative relation within the agents in any coalition  $C \in \pi$ .*

We are ready to show that any optimal mechanism is NOM. To this end, we will make sure that both the conditions of Definition 2 are satisfied.

*Worst-case)* If  $i$  truthfully reports  $w_i$ , by Lemma 3, no optimal outcome can guarantee her a negative utility as no agent in  $\delta^-(w_i)$  can be put in her coalition. Consider now any  $d_i$ , we show there exists  $\mathbf{d}_{-i}$  such that  $i$  ends up in a singleton coalition. Indeed, by Lemma 3, it suffices to set  $d_j(i) = -x$ , for each  $j \neq i$ . In conclusion, the worst-case scenario by truthfully reporting is to be assigned to the singleton coalition, and this cannot be avoided by misreporting the preferences. Therefore, condition 1 of Definition 2 is satisfied.

*Best-case)* In this case, by truthfully reporting  $w_i$ , agent  $i$  will be assigned to the coalition  $\Delta_i^+$  whenever all agents in  $\delta^+(w_i)$  positively value only agents in  $\Delta_i^+$  and any other declaration is set to  $-x$ . Since no coalition guarantees  $i$  a higher utility, condition 2 of Definition 2 is satisfied.

**Case  $x = 2n - 3$ : there exists and optimal and NOM mechanism.** Let us notice that if  $x = 2n - 3$ , Lemma 3 no longer holds true. Suppose indeed that  $C = \mathcal{N}$  and there exists a unique negative relation within the agents in  $C$ . Let  $j$  be an agent incident to the negative relation. In this case, both the outcomes  $\pi = \{\mathcal{N}\}$  and  $\pi' = \{\mathcal{N} \setminus \{j\}, \{j\}\}$  are optimal. However, from a close inspection of the proof of Lemma 3, it turns out that it still holds as long as the optimal mechanism breaks ties in favor of  $\pi'$  rather than  $\pi$ . Hence, we

can derive the following fact: for any  $\pi$  output by *this* optimal mechanism, there is no negative relation within the agents in any coalition  $C \in \pi$ . Then, by replacing Lemma 3 with the aforementioned fact in the proof for  $x < 2n - 3$ , we achieve that this optimal mechanism is NOM for  $x = 2n - 3$ .

Although we just proved that a NOM and optimal mechanism exists, it is natural to wonder if all optimal mechanisms are NOM. Unfortunately, this is not the case.

**Example 2.** *Let  $x = 2n - 3$ . Assume that in case of ties between the outcomes  $\pi = \{\mathcal{N}\}$  and  $\pi' = \{\mathcal{N} \setminus \{j\}, \{j\}\}$  are optimal, for some  $j \in \mathcal{N}$ , an optimal mechanism  $\mathcal{M}$  returns  $\pi$  rather than  $\pi'$ .*

*Consider an instance where there are three agents, namely, 1, 2, 3, and  $w_{12} = -x$  while  $w_{13} = 1$ . If 1 truthfully reports  $w_1$ , then, due to the aforementioned tie-breaking rule, the mechanism may return the grand coalition as the outcome of the game, which guarantees 1 a negative utility. Notice that this may not necessarily be the worst outcome for 1, but we can infer that, in the worst case, 1 gets a negative utility. However, if 1 reports  $d_{12} = d_{13} = -x$ , in every optimal partition,  $i$  will be put in a singleton and obtains a utility equal to 0, violating Condition 2 of Definition 2.*

*In conclusion, not every optimal mechanism is NOM.*

## 6 Conclusions and Future Work

We explored the design of mechanisms for wide classes of succinctly representable HGs that would be robust at least against manipulations of bounded rational agents and return coalition structures of good quality. We followed recent literature and considered NOM mechanisms, and we proved that these mechanisms can always guarantee non-manipulability (by agents lacking contingent reasoning) and optimality, at least in the continuous case. We also proved that if one requires computational efficiency, we can compute non-manipulable outcomes that approach the best possible approximation achievable in polynomial time. We also consider the discrete values showing a preliminary characterization of optimal and NOM mechanisms.

From a technical viewpoint, it would be clearly interesting to address the questions left open in this work: to close the gap between the best approximation achievable by a NOM mechanism in FHGs, and to extend the optimality characterization to other discrete variants of ASHG and FHGs.

Another intriguing future challenge is to understand to what extent the results and techniques defined herein can be extended beyond ASHG and FHG, either to other classes of hedonic games, such as B-games [Cechlárová and Hajduková, 2003], W-games [Cechlárová and Hajduková, 2004], or distance HGs [Flammini *et al.*, 2021b], or even to generalizations of hedonic games, such as the group activity selection problem [Darmann *et al.*, 2012]. Moreover, considering the design of mechanisms that are not manipulable by bounded rational agents beyond NOM represents an important future work. Indeed, as discussed above, NOM considers a specific way, for an agent lacking contingent reasoning, to aggregate the possible realizations of the game. Hence, it is natural to ask what may happen with other “simple” aggregation rules.

## Acknowledgments

This work was supported by: the PNRR MIUR project FAIR – Future AI Research (PE00000013), Spoke 9 – Green-aware AI; “INdAM – GNCS Project”, code CUP\_E53C24001950001.

## References

- [Amanatidis *et al.*, 2017] Georgios Amanatidis, Georgios Birmpas, George Christodoulou, and Evangelos Markakis. Truthful allocation mechanisms without payments: Characterization and implications on fairness. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 545–562, 2017.
- [Archbold *et al.*, 2023a] Thomas Archbold, Bart de Keijzer, and Carmine Ventre. Non-Obvious Manipulability for Single-Parameter Agents and Bilateral Trade. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 2107–2115, 2023.
- [Archbold *et al.*, 2023b] Thomas Archbold, Bart De Keijzer, and Carmine Ventre. Non-obvious manipulability in extensive-form mechanisms: the revelation principle for single-parameter agents. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2503–2510, 2023.
- [Archbold *et al.*, 2024] Thomas Archbold, Bart de Keijzer, and Carmine Ventre. Willy Wonka Mechanisms. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 78–86, 2024.
- [Arribillaga and Bonifacio, 2024] R. Pablo Arribillaga and Agustín G Bonifacio. Obvious manipulations of tops-only voting rules. *Games and Economic Behavior*, 143:12–24, 2024.
- [Aziz and Lam, 2021] Haris Aziz and Alexander Lam. Obvious manipulability of voting rules. In *Algorithmic Decision Theory: 7th International Conference, ADT 2021, Toulouse, France, November 3–5, 2021, Proceedings 7*, pages 179–193. Springer, 2021.
- [Aziz *et al.*, 2015] Haris Aziz, Serge Gaspers, Joachim Gudmundsson, Julián Mestre, and Hanjo Täubig. Welfare Maximization in Fractional Hedonic Games. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25–31, 2015*, pages 461–467. AAAI Press, 2015.
- [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 (TEAC)*, 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(1):135–153, 2001.
- [Bloch and Diamantoudi, 2011] Francis Bloch and Effrosyni Diamantoudi. Noncooperative formation of coalitions in hedonic games. *International Journal of Game Theory*, 40(2):263–280, 2011.
- [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.*, 2018] Florian Brandl, Felix Brandt, Manuel Eberl, and Christian Geist. Proving the incompatibility of efficiency and strategyproofness via SMT solving. *Journal of the ACM (JACM)*, 65(2):1–28, 2018.
- [Cechlárová and Hajduková, 2003] Katarína Cechlárová and Jana Hajduková. Computational complexity of stable partitions with B-preferences. *International Journal of Game Theory*, 31(3):353–364, 2003.
- [Cechlárová and Hajduková, 2004] Katarína Cechlárová and Jana Hajduková. Stable partitions with W-preferences. *Discrete Applied Mathematics*, 138(3):333–347, 2004.
- [Charness and Levin, 2009] Gary Charness and Dan Levin. The origin of the winner’s curse: a laboratory study. *American Economic Journal: Microeconomics*, 1(1):207–236, 2009.
- [Darmann *et al.*, 2012] Andreas Darmann, Edith Elkind, Sascha Kurz, Jérôme Lang, Joachim Schauer, and Gerhard Woeginger. Group activity selection problem. In *Internet and Network Economics: 8th International Workshop, WINE 2012, Liverpool, UK, December 10–12, 2012. Proceedings 8*, pages 156–169. Springer, 2012.
- [Dimitrov and Sung, 2004] Dinko Dimitrov and Shao Chin Sung. Enemies and Friends in Hedonic Games: Individual Deviations, Stability and Manipulation. *SSRN Electronic Journal*, 02 2004.
- [Dimitrov *et al.*, 2006] Dinko Dimitrov, Peter Borm, Ruud Hendrickx, and Shao Chin Sung. Simple priorities and core stability in hedonic games. *Social Choice and Welfare*, 26:421–433, 2006.
- [Dreze and Greenberg, 1980] Jacques H. Dreze and Joseph Greenberg. Hedonic coalitions: Optimality and stability. *Econometrica: Journal of the Econometric Society*, pages 987–1003, 1980.
- [Dur *et al.*, 2018] Umut Dur, Robert G. Hammond, and Thayer Morrill. Identifying the harm of manipulable school-choice mechanisms. *American Economic Journal: Economic Policy*, 10(1):187–213, 2018.
- [Elkind and Wooldridge, 2009] Edith Elkind and Michael J. Wooldridge. Hedonic coalition nets. In *8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), Budapest, Hungary, May 10–15, 2009, Volume 1*, pages 417–424. IFAAMAS, 2009.
- [Esponda and Vespa, 2014] Ignacio Esponda and Emanuel Vespa. Hypothetical thinking and information extraction in the laboratory. *American Economic Journal: Microeconomics*, 6(4):180–202, 2014.
- [Feldman *et al.*, 2015] Moran Feldman, Liane Lewin-Eytan, and Joseph (Seffi) Naor. Hedonic Clustering Games. *ACM Trans. Parallel Comput.*, 2(1):4:1–4:48, may 2015.



- [Flammini *et al.*, 2021a] Michele Flammini, Bojana Kodric, Gianpiero Monaco, and Qiang Zhang. Strategyproof mechanisms for additively separable and fractional hedonic games. *Journal of Artificial Intelligence Research*, 70:1253–1279, 2021.
- [Flammini *et al.*, 2021b] Michele Flammini, Bojana Kodric, Martin Olsen, and Giovanna Varricchio. Distance Hedonic Games. In *SOFSEM 2021: Theory and Practice of Computer Science - 47th International Conference on Current Trends in Theory and Practice of Computer Science, SOFSEM 2021, Bolzano-Bozen, Italy, January 25-29, 2021, Proceedings*, volume 12607 of *Lecture Notes in Computer Science*, pages 159–174. Springer, 2021.
- [Flammini *et al.*, 2022] Michele Flammini, Bojana Kodric, and Giovanna Varricchio. Strategyproof mechanisms for Friends and Enemies Games. *Artif. Intell.*, 302:103610, 2022.
- [Flammini *et al.*, 2025] Michele Flammini, Maria Fomenko, and Giovanna Varricchio. Non-obvious Manipulability in Hedonic Games with Friends Appreciation Preferences. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 767–775. ACM, 2025.
- [Gairing and Savani, 2010] Martin Gairing and Rahul Savani. Computing stable outcomes in hedonic games. In *International Symposium on Algorithmic Game Theory*, pages 174–185. Springer, 2010.
- [Hajduková, 2006] Jana Hajduková. Coalition formation games: A survey. *International Game Theory Review*, 8(04):613–641, 2006.
- [Igarashi and Elkind, 2016] Ayumi Igarashi and Edith Elkind. Hedonic games with graph-restricted communication. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 242–250. International Foundation for Autonomous Agents and Multiagent Systems, 2016.
- [Klaus *et al.*, 2023] Bettina-Elisabeth Klaus, Flip Klijn, and Seçkin Özbilen. Core Stability and Strategy-Proofness in Hedonic Coalition Formation Problems with Friend-Oriented Preferences. *Available at SSRN 4677609*, 2023.
- [Ngangoué and Weizsäcker, 2021] M. Kathleen Ngangoué and Georg Weizsäcker. Learning from unrealized versus realized prices. *American Economic Journal: Microeconomics*, 13(2):174–201, 2021.
- [Ortega and Segal-Halevi, 2022] Josué Ortega and Erel Segal-Halevi. Obvious manipulations in cake-cutting. *Social Choice and Welfare*, 59(4):969–988, 2022.
- [Özyurt and Sanver, 2009] Selçuk Özyurt and M Remzi Sanver. A general impossibility result on strategy-proof social choice hyperfunctions. *Games and economic behavior*, 66(2):880–892, 2009.
- [Pathak and Sönmez, 2008] Parag A. Pathak and Tayfun Sönmez. Leveling the playing field: Sincere and sophisticated players in the boston mechanism. *American Economic Review*, 98(4):1636–1652, 2008.
- [Psomas and Verma, 2022] Alexandros Psomas and Paritosh Verma. Fair and efficient allocations without obvious manipulations. *Advances in Neural Information Processing Systems*, 35:13342–13354, 2022.
- [Rodríguez-Álvarez, 2009] Carmelo Rodríguez-Álvarez. Strategy-proof coalition formation. *International Journal of Game Theory*, 38:431–452, 2009.
- [Soifer, 2009] Alexander Soifer. *The mathematical coloring book: Mathematics of coloring and the colorful life of its creators*. Springer, 2009.
- [Troyan and Morrill, 2020] Peter Troyan and Thayer Morrill. Obvious manipulations. *Journal of Economic Theory*, 185:104970, 2020.
- [Varricchio, 2023] Giovanna Varricchio. On Approximate Strategyproof Mechanisms for Hedonic Games and the Group Activity Selection Problem. In *Proceedings of IPS*, volume 3585 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2023.