Non-obvious Manipulability in Hedonic Games with Friends Appreciation Preferences

Michele Flammini
Gran Sasso Science Institute
L'Aquila, Italy
michele.flammini@gssi.it

Maria Fomenko
Gran Sasso Science Institute
L'Aquila, Italy
maria.fomenko@gssi.it

Giovanna Varricchio
University of Calabria
Rende, Italy
giovanna.varricchio@unical.it

ABSTRACT

In this paper, we study non-obvious manipulability (NOM), a relaxed form of strategyproofness, in the context of Hedonic Games (HGs) with Friends Appreciation (FA) preferences. In HGs, the aim is to partition agents into coalitions according to their preferences which solely depend on the coalition they are assigned to. Under FA preferences, agents consider any other agent either a friend or an enemy, preferring coalitions with more friends and, in case of ties, the ones with fewer enemies. Our goal is to design mechanisms that prevent manipulations while optimizing social welfare.

Prior research established that computing a welfare maximizing (optimum) partition for FA preferences is not strategyproof, and the best-known approximation to the optimum subject to strategyproofness is linear in the number of agents. In this work, we explore NOM to improve approximation results. We first prove the existence of a NOM mechanism that always outputs the optimum; however, we also demonstrate that the computation of an optimal partition is NP-hard. To address this complexity, we focus on approximation mechanisms and propose a NOM mechanism guaranteeing a $(4 + o(1))$-approximation in polynomial time.

Finally, we briefly discuss NOM in the case of Enemies Aversion (EA) preferences, the counterpart of FA, where agents give priority to coalitions with fewer enemies and show that no mechanism computing the optimum can be NOM.

KEYWORDS

Hedonic Games; Strategyproofness; Non-obvious Manipulability

ACM Reference Format:
Michele Flammini, Maria Fomenko, and Giovanna Varricchio. 2025. Non-obvious Manipulability in Hedonic Games with Friends Appreciation Preferences. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

1 INTRODUCTION

Hedonic Games (HGs) [13] offer a game-theoretic framework for understanding the coalition formation of selfish agents and have been extensively studied in the literature (e.g.,[2, 3, 5, 8, 14, 15, 21]). In such games, the objective is to partition a set of agents into disjoint coalitions, with each agent's satisfaction determined solely by the members of her coalition. Depending on the nature of the agents' preferences, several HGs classes arise that may capture various social interactions between agents. For example, in *additively separable* HGs (ASHGs) [23], agents evaluate coalitions by summing up the values they assign to every other participant; in HGs with *friends appreciation* (FA) preferences [12], each agent splits the others into friends and enemies and prefers coalitions with more friends; in case of ties, she favors the ones with less enemies.

Most of the existing literature on HGs has put attention on the existence and computation of several stability concepts based on either individual [7, 16, 18, 21] or group deviations [5, 8, 15, 21, 24]. However, a recent stream of research is focusing on designing *strategyproof* (SP) mechanisms. Such mechanisms prevent agents from manipulating the outcome by misrepresenting their preferences while ensuring desirable properties like stability or a reasonable approximation to the maximum *social welfare* – the sum of the agents' utilities in the outcome. Unfortunately, achieving strategyproofness with good social welfare guarantees is challenging: even in the simple case of FA preferences the best-known SP mechanism guarantees an approximation linear in the number of agents [19].

Although strategyproofness has been widely studied in several game-theoretic settings, it turned out to be often incompatible with other desirable properties or even impossible to achieve [1, 9, 17, 27]. Moreover, there exist mechanisms that are not strategyproof in a strict sense, but, in order to successfully manipulate, an agent has to possess the knowledge of others' strategies and deeply understand the underlying mechanics. Otherwise, she might end up with an outcome that is even worse than the one she attempted to avoid. However, the ability of a cognitively limited agent to satisfy this requirement seems unrealistic, which has led to the notion of *non-obvious manipulability* (NOM) introduced to distinguish the mechanisms that can be easily manipulated from the ones that are unlikely to be manipulated in practice [32].

Our Contribution. We initiate the study of NOM in the context of HGs focusing on FA preferences. We aim at improving upon the performances, in terms of the social welfare guarantee, of SP mechanisms in this setting. To this end, we begin by analyzing the structure of optimal outcomes and give a deeper understanding of how such outcomes look like for some interesting instances. We specifically focus on some structures of friendship relationships which turn out to be very useful for providing a picture of socially optimal outcomes. This enables us to show that there always exists a NOM mechanism computing a social welfare maximizing partition (Theorem 1). Unfortunately, we also show that finding such an outcome is NP-hard (Theorem 2). We, therefore, propose a NOM polynomial-time mechanism with $(4 + o(1))$-approximation (Theorem 3). It provides a significant improvement over the existing strategyproof mechanism, and, besides NOM, constitutes the

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

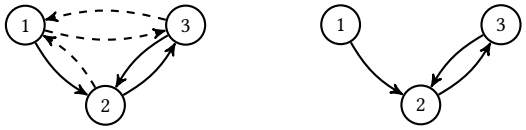
first constant approximation for the problem of maximizing the utilitarian welfare for FA preferences. Finally, we investigate NOM for *enemies aversion* (EA) preferences, the natural counterpart of FA where agents give priority to coalitions with fewer enemies, and show that no optimal mechanism is NOM (Theorem 4). This shows that NOM, albeit it might be considered a weak notion, is not always compatible with optimality. Due to the space limits, missing proofs are deferred to the full paper.

Related Work. Hedonic games with FA and EA preferences have been widely studied and further extended to capture more involved social contexts [6, 11, 25, 30]. In this stream of research, a systematic analysis of stable outcomes has been provided [10]. In addition to stability, strategyproofness has also been considered: in [12] the authors show that for FA and EA preferences SP is compatible with (weak) core stability. However, such solutions are hard to compute in the case of EA preferences [11]. For more general friends-oriented preferences [26] and ASHGs [29], SP mechanisms guaranteeing stability have been investigated. Some recent studies, instead of seeking stability, have concentrated on strategyproof mechanisms approximating the maximum social welfare [33]. For ASHGs in general, even when the agents' utilities are bounded, it was proven that a non-manipulable algorithm with a bounded approximation ratio cannot exist [17]; the authors also provide bounded, but non-constant SP mechanisms for very restricted settings. Similarly, in the FA model, a deterministic mechanism with the approximation ratio linear in the number of agents and a randomized one with a constant approximation ratio have been provided [19]. Also, in the case of EA preferences, the best-known polynomial algorithm achieves a linear approximation in the number of agents, while a constant approximation ratio is possible when time complexity is not a concern [19], and this result has been proven to be asymptotically tight. Some attempts in achieving SP and bounded approximation have been made also for a proper superclass of HGs, namely, the group activity selection problem [20].

In contrast to SP, in the past few years, non-obvious manipulability has been introduced [32]. This notion turned out to be a relaxation good enough to circumvent the inherent impossibility results of strategyprofness. In voting theory, non-obvious manipulability allows to bypass of a famous strong negative result stating the non-existence of a voting rule for more than two alternatives which is at the same time strategyproof and not dictatorial [4]. In the assignment problem, under a minor restriction the whole class of rank-minimising mechanisms, which directly optimize an objective natural for this problem, turns out to be NOM [31]. In the problem of fairly allocating indivisible goods, replacing strategyproofness with non-obvious manipulability allows the design of a Pareto-efficient and non-dictatorial mechanism as well as a mechanism that guarantees envy-freeness up to one item [28]. Since in HGs strategyproof mechanisms often fail to approximate the maximum social welfare with a constant ratio or turn out to be ineffective in the computational sense, this provides us with additional motivation to study NOM mechanisms.

2 PRELIMINARIES

Hedonic Games and Friends Appreciation preferences. In the classical framework of HGs, we are given a set of n agents, denoted by $\mathcal{N} = \{1, \dots, n\}$, and the goal is to partition them into disjoint coalitions. In other words, we aim at creating a disjoint partition $\pi = \{C_1, \dots, C_m\}$ such that $\cup_{h=1}^{m} C_h = \mathcal{N}$ and $C_h \cap C_k = \emptyset$ for $h \neq k$. Such a partition is also called an *outcome* or a *coalition structure*. The *grand coalition* $\mathcal{GC}$ is a partition consisting of exactly one coalition containing all the agents, while a *singleton coalition* is any coalition of size 1. We denote by Π the set of all possible outcomes of the game, i.e., all possible partitions of the agents, and by $\pi(i)$ the coalition that agent i belongs to in a given outcome $\pi \in \Pi$.



(a) An FA instance $\mathcal{I}$. **(b) The friendship graph $G^f(\mathcal{I})$.**

Figure 1: An FA instance and the corresponding G^f. Solid (resp. dashed) edges represent friend (resp. enemy) relations.

In HGs, the agents evaluate an outcome on the sole basis of the coalition they belong to and not on how the others aggregate. As a result, each agent i has a preference relation $\succeq_i$ over $\mathcal{N}_i$, where $\mathcal{N}_i$ is the family of subsets of $\mathcal{N}$ containing i. Given $X, Y \in \mathcal{N}_i$, we say that agent i prefers, or equally prefers, X to Y whenever $X \succeq_i Y$.

In HGs with *friends appreciation* (FA) preferences, each agent i partitions the other agents into a set of friends F_i and a set of enemies E_i, with $F_i \cup E_i = \mathcal{N} \setminus \{i\}$ and $F_i \cap E_i = \emptyset$. The preferences of i among coalitions in $\mathcal{N}_i$ are as follows: $X \succeq_i Y$ if and only if

$$|X \cap F_i| > |Y \cap F_i| \text{ or}$$
$$|X \cap F_i| = |Y \cap F_i| \text{ and } |X \cap E_i| \leq |Y \cap E_i|.$$

In other words, a coalition is preferred over another one if it contains a higher number of friends; if the number of friends is the same, the coalition with fewer enemies is preferred.

Example 1. Let us describe a simple instance with friends appreciation preferences: Let $\mathcal{N} = \{1, 2, 3\}$ be the set of agents, and let $F_1 = \{2\}$, $F_2 = \{3\}$, $F_3 = \{2\}$ and $E_1 = \{3\}$, $E_2 = \{1\}$, $E_3 = \{1\}$ be the agents' sets of friends and enemies, respectively. This instance is depicted in Figure 1a, where a directed edge from agent i to agent j represents i's opinion of j; solid edges and dashed edges represent friend and enemy relations, respectively.

For our convenience, we shall denote by $F_i^{-1} = \{j \in \mathcal{N} \mid i \in F_j\}$, that is, the set of agents considering i a friend.

FA are a proper subclass of ASHGs, where each agent i has a value $v_i(j)$ for every other agent j and her utility for being in a given coalition $C \in \mathcal{N}_i$ is $u_i(C) = \sum_{j \in C \setminus \{i\}} v_i(j)$. Specifically, to comply with the FA preferences, u_i can be defined as follows:

$$v_i(j) = 1, \text{ if } j \in F_i, \qquad \text{and} \qquad v_i(j) = -\frac{1}{n}, \text{ if } j \in E_i.$$

For every agent the sum of the absolute values of all enemies never exceeds the value of one friend, therefore, FA preferences are correctly encoded. These valuation functions were already assumed in [11, 17]. Since the utility of an agent depends only on the coalition she belongs to, we might write $u_i(\pi)$ to denote $u_i(\pi(i))$.

Example 2. Let us consider the instance described in Example 1 and the partition $\pi = \{\{1,2,3\}\}$. Then, $u_1(\pi) = v_1(2) + v_1(3) = 1 - \frac{1}{3} = \frac{2}{3}$. Similarly, since the utility of an agent depends only on the number of friends/enemies in her coalition, $u_2(\pi) = u_3(\pi) = \frac{2}{3}$.

An FA instance $\mathcal{I}$ is given by a set of agents $\mathcal{N}$ and a set of friends F_i, for each $i \in \mathcal{N}$. Alternatively, $\mathcal{I} = (\mathcal{N}, \{v_i\}_{i\in\mathcal{N}})$, where v_i is the valuation of i for the other agents representing her FA preferences. For simplicity, we might also write $\mathcal{I} = (\{v_i\}_{i\in\mathcal{N}})$.

Social Welfare and Optimum. One of the challenges in HGs is to maximize the overall happiness of the agents measured by the *social welfare* (SW). Specifically, in an HG instance $\mathcal{I}$ the social welfare of a partition π is given by $\mathrm{SW}^{\mathcal{I}}(\pi) = \sum_{i\in\mathcal{N}} u_i(\pi)$.

When the instance is clear from the context we simply write SW. We call *social optimum*, or simply *optimum*, any outcome OPT in $\arg\max_{\pi\in\Pi} \mathrm{SW}(\pi)$ and denote by opt the value SW(OPT). When considering a coalition C, to denote $\sum_{i\in C} u_i(C)$ we write $\mathrm{SW}(C)$.

Graph Representation. A very convenient representation of an FA instance $\mathcal{I}$ is by means of a directed and unweighted graph where the agents of the instance are the vertices. With E_i being $\mathcal{N} \setminus \{F_i \cup \{i\}\}$, we represent only friendship relationships through directed edges: if $\{i,j\}$ is an edge of this graph, it means $j \in F_i$; if such an edge does not exist, we have $j \in E_i$. We call this graph the *friendship graph* of $\mathcal{I}$, and we denote it by $G^f(\mathcal{I}) = (\mathcal{N}, F)$, where $F = \{\{i,j\} \mid j \in F_i\}$. If the instance is clear from the context, we simply write G^f. The friendship graph of the instance described in Example 1 is shown in Figure 1b. Moreover, we denote by $N(i)$, for $i \in \mathcal{N}$, the weakly connected neighborhood of i, that is, $N(i) = F_i \cup F_i^{-1}$. We denote by $\delta(i)$ the size of $N(i)$. We further extend the definition of weekly connected neighborhood to a subset of agents $X \subseteq \mathcal{N}$, specifically $N(X) = \cup_{i\in X} N(i)$.

Strategyproofness and Non-obvious Manipulability. The sets F_i and E_i might be private information of the agent i; therefore, to compute the outcome we need to receive this information from the agents. Let us denote by $\mathbf{d} = (d_1, \ldots, d_n)$ the agents' *declarations* vector, where d_i contains the information related to agent i. We assume direct revelation, and hence $d_i(j) \in \{1, -\frac{1}{n}\}$ represents the value i declared for an agent j. We denote by $\mathcal{D}$ the space of feasible declarations $\mathbf{d}$. For our convenience, we denote by $\mathbf{d}_{-i}$ the agents' declarations except the one of i, by $\mathcal{D}_{-i}$ the set of all feasible $\mathbf{d}_{-i}$, and by $\mathcal{D}_i$ the feasible declarations for i.

In this setting, the natural challenge is to design algorithms, a.k.a. *mechanisms*, inducing truthful behavior of the agents. We shall denote by $\mathcal{M}$ a mechanism and by $\mathcal{M}(\mathbf{d})$ the output of the mechanism – a partition upon the declaration $\mathbf{d}$ of the agents.

The agents might be strategic, which means, an agent i could declare $d_i \neq t_i$, where $t_i \in \mathcal{D}_i$ is the real information of agent i, also called her *real type*. For this reason, the design of mechanisms preventing manipulations is fundamental. The most desirable characteristic for such kind of mechanisms is *strategyproofness*.

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

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

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

In turn, a mechanism is said to be *manipulable* if there exists an agent i, a real type t_i and declarations $\mathbf{d}_{-i}$ and $d_i \neq t_i$ such that Equation (1) does not hold. Then, such d_i is called a *manipulation*.

Since SP mechanisms may be quite inefficient w.r.t. the truthful opt, we aim to understand if mechanisms satisfying milder conditions lead to more efficient outcomes. Considering that i might be unaware of which are the declarations $\mathbf{d}_{-i}$ of the other agents, she could not be able to determine a manipulation without knowing $\mathbf{d}_{-i}$. Thus, we next consider a relaxation of SP where an agent i decides to misreport her true values only if it is clearly profitable for her. Given a mechanism $\mathcal{M}$, let us denote by $\Pi_i(d_i, \mathcal{M}) = \{\mathcal{M}(d_i, \mathbf{d}_{-i}) \mid \mathbf{d}_{-i} \in \mathcal{D}_{-i}\}$, the space of possible outcomes of $\mathcal{M}$ given the declaration d_i of i. Notice the space $\Pi_i(d_i, \mathcal{M})$ is finite.

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 t_i, and any other declaration d_i the following hold true:

Condition 1: $\max_{\pi\in\Pi_i(t_i,\mathcal{M})} u_i(\pi) \geq \max_{\pi\in\Pi_i(d_i,\mathcal{M})} u_i(\pi)$

Condition 2: $\min_{\pi\in\Pi_i(t_i,\mathcal{M})} u_i(\pi) \geq \min_{\pi\in\Pi_i(d_i,\mathcal{M})} u_i(\pi)$

If there exist i, t_i, and d_i such that Condition 1 or 2 is violated, then, $\mathcal{M}$ is *obviously manipulable* and d_i is an *obvious manipulation*.

In other words, a mechanism $\mathcal{M}$ is NOM if for every agent i neither the best nor the worst possible outcome can be improved in terms of i's utility by manipulating, i.e., declaring some d_i instead of t_i (worst/best outcomes are always determined according to i's true preferences). In contrast, the strategyproofness of a mechanism $\mathcal{M}$ ensures that for every $\mathbf{d}_{-i}$, including the ones inducing the best/worst case outcome of $\Pi_i(t_i, \mathcal{M})$, is not strictly convenient to misreport; therefore, SP $\Rightarrow$ NOM.

In what follows, we always denote by t_i the real type of i and by $e_i = |E_i|$ and $f_i = |F_i|$, where E_i and F_i are the truthful set of friends and enemies of i, respectively.

2.1 Preliminary Results on Optimal Outcomes

In this section, we discuss in detail some useful properties of optimal outcomes putting particular attention on specific graph structures for the friendship graph. Let us start by observing that to compute the social welfare of a coalition it suffices to know its size and the number of friendship relationships within the coalition.

Lemma 1 (From [19])**.** *For any $C \subseteq \mathcal{N}$ of size c and containing f_C friendship relations,* $\mathrm{SW}(C) = f_C \cdot \left(1 + \frac{1}{n}\right) - \frac{c(c-1)}{n}$.

Example 3. Consider, for example, an FA instance where G^f is a star whose edges are directed from its center i towards the leaves, that is, $F = \{\{i,j\} \mid j \in \mathcal{N} \setminus \{i\}\}$. For such an instance, if we put the agents together in the grand coalition, we have $\mathrm{SW}(\mathcal{GC}) = \frac{n-1}{n}$. Clearly, in any optimum π^*, any node that is not in the same coalition as i must be in a singleton coalition: otherwise, the SW of its coalition would be negative. Let C be the coalition of i in π^*, $\mathrm{SW}(\pi^*) = \mathrm{SW}(C) = (c-1)\cdot\left(1 + \frac{1}{n} - \frac{c}{n}\right)$, which is maximized at $c = \frac{n+2}{2}$, for even n, and at $c = \frac{n+2}{2} \pm \frac{1}{2}$, otherwise. Note that optimality does not depend on the edges direction.

Example 3 shows that when the graph is particularly sparse it is more convenient to split weakly connected components rather than put all agents with all their friends. In turn, when there exists a cluster of nodes $C \subset \mathcal{N}$ such that C is a bidirectional clique in G^f, whose nodes are weakly connected only to another node $i \in \mathcal{N} \setminus C$, it is never convenient to split the agents in C. We call C an *almost isolated clique* with the *hinge node* i. We next formalize how almost isolated cliques place in an optimum outcome.

Lemma 2. *If C is an almost isolated clique in G^f with the hinge node i, for any optimal partition π^* there exists $C' \in \pi^*$ such that $C \subseteq C'$. Furthermore, if $i \notin C'$ then $C' = C$.*

Next, we consider more involved structures for G^f and explain how their optimal outcomes look like.

Definition 3 (Octopus Graph)**.** Given an agent i and $H \subseteq \mathcal{N} \setminus \{i\}$, $G^f = (\mathcal{N}, F)$ is an i-centered *octopus graph* with the head H if:

- H is a bidirectional clique in G^f;
- for each $j \in H$, $\{j, i\} \in F$;
- for each $j \in \mathcal{N} \setminus i$ and $k \in \mathcal{N} \setminus (\{i\} \cup H)$, none of $\{j, k\}$, $\{k, j\}$, $\{k, i\}$ belongs to F while $\{i, j\}$ may belong to F.

A picture of an i-centered octopus graph is given in Figure 2a.

Lemma 3. *Let G^f be an i-centered octopus graph with the head H. If $|H| \geq \lceil \frac{n}{2} \rceil$, there exists a unique optimum consisting of the coalition $H \cup \{i\}$ and remaining agents put in singleton coalitions.*

Sketch. Let us start by noticing that H is an almost isolated clique with the hinge i. By Lemma 2, in the social optimum, all agents from H will end up in the same coalition. Moreover, for each agent $k \in \mathcal{N} \setminus (\{i\} \cup H)$, k can be weakly connected only to i, so, if k is not in the same coalition as i, then k forms a singleton coalition. This leaves us with three possible types of optimal partition:

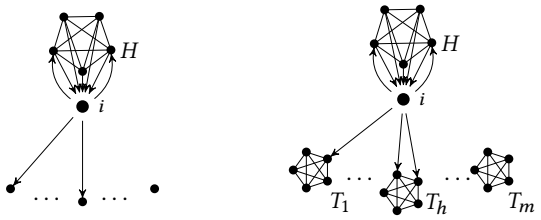
- π^1 where H and i are in the same coalition, while all remaining agents are in singletons;
- π^2 where agents from H form a coalition, while i is in a different coalition together with some $C_1 \subseteq \mathcal{N} \setminus (\{i\} \cup H)$, and remaining agents are in singletons;
- π^3 where H, i and some $C_2 \subseteq \mathcal{N} \setminus (\{i\} \cup H)$ form one coalition, while all other agents are in singletons.

Let us compare π^1 and π^2. The number of positive relationships between H and i is at least $|H|$ while the positive connections between i and C_1 are at most $|C_1|$. Since $|C_1| \leq n - |H| - 1 \leq n - \lceil \frac{n}{2} \rceil - 1 < |H|$, it is strictly more convenient to put i in coalition with H rather than with C_1, showing $\mathrm{SW}(\pi^1) > \mathrm{SW}(\pi^2)$.

Let us compare π^1 and π^3. Consider the coalition $H \cup C_2 \cup \{i\}$. The friendship relationships between $H \cup \{i\}$ and C_2 are only the ones in $F_i \cap C_2$, which are at most $|C_2|$. So, if we remove and split C_2 into singletons, the loss in social welfare will be of at most $|C_2| \cdot \left(1 + \frac{1}{n}\right) - \frac{2 \cdot (|H|+1) \cdot |C_2|}{n} \leq |C_2| \cdot \frac{n+1-2 \cdot \frac{n}{2}-2}{n} < 0$, where the first inequality holds as $|H| \geq \frac{n}{2}$. Therefore, $\mathrm{SW}(\pi^1) > \mathrm{SW}(\pi^3)$. □

We now further generalize the definition of octopus graph.

Definition 4 (Generalized Octopus Graph)**.** Given an agent $i \in \mathcal{N}$ and $\{H, T_1, \ldots, T_m\}$, a disjoint partition of $\mathcal{N} \setminus \{i\}$; $G^f = (\mathcal{N}, F)$ is an i-centered *generalized octopus graph* with the head H and the tentacles $T_1, \ldots, T_m$ if, for each $l \in [m]$:



(a) Octopus graph. **(b) Generalized octopus graph.**

Figure 2: Octopus graph structures having center i. Undirected edges represent bidirectional edges in G^f.

- H and T_l are bidirectional cliques in G^f;
- for each $j \in H$, $\{j, i\} \in F$;
- for each $j \in \mathcal{N} \setminus \{T_l \cup \{i\}\}$ and $k \in T_l$, none of $\{j, k\}$, $\{k, j\}$, $\{k, i\}$ belongs to F while $\{i, j\}$ may belong to F.

Given a node i, we denote by $gOct(i)$ the set of all possible i-centered generalized octopus graphs. In Figure 2b, we draw an example of an i-centered generalized octopus graph. Let us note that an octopus graph is a generalized octopus graph having all tentacles of size 1. Furthermore, if $G^f(d_i, \mathbf{d}_{-i}) \in gOct(i)$, for any declaration $d_i' \in \mathcal{D}_i$, $G^f(d_i', \mathbf{d}_{-i}) \in gOct(i)$; in fact, in the definition there is no constraint on how the set of friends of i should be.

Lemma 4. *Let G^f be a generalized i-centered octopus graph with head H and tentacles $T_1, \ldots, T_m$. If $\pi^* = \{H \cup \{i\}, T_1, \ldots, T_m\}$ is an optimum outcome, then $|H| \geq \max_{l \in [m]} \frac{|F_i \cap T_l|}{|T_l|} \cdot \frac{n+1}{2} - 1$.*

Proof. Let us set $g_l = |F_i \cap T_l|$. Consider the partition π obtained from π^* by merging the coalitions $H \cup \{i\}$ and T_l. Since π^* is optimum, $\mathrm{SW}(\pi) - \mathrm{SW}(\pi^*) = g_l - \frac{2|T_l| - g_l}{n} - \frac{2|T_l| \cdot |H|}{n} \leq 0$. Therefore, $|H| \geq \frac{g_l(n+1)}{2|T_l|} - 1$, and hence it holds for the maximum. □

Next, we provide an interesting connection between the optimum for arbitrary instances and the one for octopus graphs.

Lemma 5. *If π^* is an optimum partition for $\mathcal{I} = (d_i, d_{-i})$, then, π^* is also optimum for $\mathcal{I}' = (d_i, d'_{-i})$, where $G^f(\mathcal{I}') \in gOct(i)$ with the head $\pi(i) \setminus \{i\}$ and the remaining coalitions of π being the tentacles.*

In a nutshell: Strengthening friendships within and enmities between the coalitions of π maintains optimality.

Proof. Let us denote by $F^{\mathcal{I}}$ and $F^{\mathcal{I}'}$ the friendship relationship in $\mathcal{I}$ and $\mathcal{I}'$, respectively. We set $p = \left|F^{\mathcal{I}'} \setminus F^{\mathcal{I}}\right|$ and $q = \left|F^{\mathcal{I}} \setminus F^{\mathcal{I}'}\right|$, which denote the number of added and removed edges when transforming $G^f(\mathcal{I})$ into $G^f(\mathcal{I}')$. Then, we have $\mathrm{SW}^{\mathcal{I}'}(\pi) - \mathrm{SW}^{\mathcal{I}}(\pi) = p \cdot \left(1 + \frac{1}{n}\right)$ since the edges we added are within coalitions and the edges we removed are between coalitions. Consider now an outcome $\pi' \neq \pi$ and let p' and q' be the number of the added and removed edges, whose endpoints are in the same coalition in π', when transforming $G^f(\mathcal{I})$ into $G^f(\mathcal{I}')$. In this case, being $p' \leq p$,

$$\mathrm{SW}^{\mathcal{I}'}(\pi') - \mathrm{SW}^{\mathcal{I}}(\pi') = p' \cdot \left(1 + \frac{1}{n}\right) - q' \cdot \left(1 + \frac{1}{n}\right) \leq p \cdot \left(1 + \frac{1}{n}\right).$$

Therefore, $\mathrm{SW}^{\mathcal{I}'}(\pi) - \mathrm{SW}^{\mathcal{I}'}(\pi') \geq \mathrm{SW}^{\mathcal{I}}(\pi) - \mathrm{SW}^{\mathcal{I}}(\pi') \geq 0$, for any $\pi' \in \Pi$, showing the optimality of π for $\mathcal{I}'$. □

3 AN OPTIMAL AND NOM MECHANISM

In [19], it has been shown that no strategyproof mechanism can have an approximation better than 2. In contrast, we next show there is a way to simultaneously guarantee optimality and NOM.

Let us first introduce our optimal mechanism.

Mechanism $\mathcal{M}_1$. It always returns an optimum partition with the smallest number of coalitions, ties among partitions with the same number of coalitions are broken arbitrarily.

Theorem 1. *Mechanism $\mathcal{M}_1$ is NOM.*

To show the theorem, we at first need to understand which are the worst/best outcomes for i in $\Pi_i(t_i, \mathcal{M}_1)$, the space of possible outcomes of $\mathcal{M}_1$ when i reports t_i. We will then compare their utility for i with the one of the worst/best outcomes in $\Pi_i(d_i, \mathcal{M}_1)$ for any other feasible d_i. Recall that $e_i = |E_i|$ and $f_i = |F_i|$ are the sizes of the truthful set of friends and enemies of i, respectively.

Lemma 6. *For any agent $i \in \mathcal{N}$, among the outcomes in $\Pi_i(t_i, \mathcal{M}_1)$:*

(1) any outcome where i is put in the coalition $\{i\} \cup F_i$ maximizes the utility of i;

(2) any outcome where i is put in a coalition with E_i and $\max\{\lceil \frac{n}{2} \rceil - e_i, 0\}$ friends minimizes the utility of i.

Proof. We first show 1. Given t_i, let us set $\mathbf{d}_{-i}$ in such a way that $\{i\} \cup F_i$ is a bidirectional clique in the corresponding friendship graph G^f while all remaining agents are isolated nodes. For such instance, there is a unique optimum and it is attained by the partition where $\{i\} \cup F_i$ is the only non-singleton coalition. Clearly, no partition can guarantee i strictly higher utility, therefore 1 follows.

We now show 2 and distinguish between $e_i \geq \lceil \frac{n}{2} \rceil$ and $e_i < \lceil \frac{n}{2} \rceil$.

If $e_i \geq \lceil \frac{n}{2} \rceil$, consider a declaration of the others $\mathbf{d}_{-i}$ such that the friendship graph $G^f(t_i, \mathbf{d}_{-i})$ is an i-centered octopus graph with the head E_i. By Lemma 3, there exists a unique optimum, which is therefore the output of $\mathcal{M}_1$, and $E_i \cup \{i\}$ is one of its coalitions. Since there exists no coalition where i gets strictly lower utility, we can conclude it is the worst possible outcome of $\mathcal{M}_1$ for i.

Assume now $e_i < \lceil \frac{n}{2} \rceil$. We first show there exists $\mathbf{d}_{-i}$ such that in $\mathcal{M}_1(t_i, \mathbf{d}_{-i})$, i is in a coalition with the agents $C \subset \mathcal{N} \setminus \{i\}$ with $f' = \lceil \frac{n}{2} \rceil - e_i$ many friends. In fact, in this case, $|C| \geq \lceil \frac{n}{2} \rceil$ and we can build as before $\mathbf{d}_{-i}$ so as the friendship graph $G^f(t_i, \mathbf{d}_{-i})$ is an i-centered octopus graph with the head C and apply Lemma 3. To conclude this is the worst outcome for i declaring t_i, we next show there is no outcome where i has strictly less friends.

Assume i is put into a coalition together with $C \subset \mathcal{N} \setminus \{i\}$ where $f' = |C \cap F_i| < \lceil \frac{n}{2} \rceil - e_i$. Note that having the same number of friends and a greater number of enemies is not possible.

Assume such $C \cup \{i\}$, having size $c = |C|$, is a coalition in the social optimum, say $\pi^* = \{C \cup \{i\}, T_1, \ldots, T_m\}$, for a game instance $(d_i, \mathbf{d}_{-i})$. Let us then select $\mathbf{d}'_{-i}$ so that $G^f(d_i, \mathbf{d}'_{-i})$ is an i-centered generalized octopus graph having head C and tentacles $T_1, \ldots, T_m$. From the proof of Lemma 5 π^* is optimum also for $(d_i, \mathbf{d}'_{-i})$. By Lemma 4, π^* is optimum only if $c > \max_l \left\{ \frac{|T_l \cap F_i|}{|T_l|} \right\} \cdot \frac{n+1}{2} - 1$. We notice that, in Lemma 4, the last inequality is $\geq$; however, the proof can be easily adjusted to show that for $\mathcal{M}_1$ the inequality turns out to be strict, as the mechanism selects an optimum with the lowest number of coalitions.

Now, $\max_l \left\{ \frac{|T_l \cap F_i|}{|T_l|} \right\} \geq \frac{\sum_{l=1}^m |T_l \cap F_i|}{\sum_{l=1}^m |T_l|} = \frac{\left|\left(\bigcup_{l \in [m]} T_l\right) \cap F_i\right|}{\left|\bigcup_{l \in [m]} T_l\right|} = \frac{|T \cap F_i|}{|T|}$, where $T = \bigcup_{l \in [m]} T_l$. This implies, $c > \frac{|T \cap F_i|}{|T|} \cdot \frac{n+1}{2} - 1$. Therefore, $c > \frac{n-1-e_i-f'}{n-c-1} \cdot \frac{n+1}{2} - 1$ as $T = \mathcal{N} \setminus \{C \cup \{i\}\}$ and $|T \cap F_i| = f_i - f' = n - 1 - e_i - f'$, and hence

$$f' > n - 1 - e_i - (c+1)(n-c-1) \cdot \frac{2}{n+1}.$$

Now, recall that $f' \leq \lceil \frac{n}{2} \rceil - e_i - 1$, thus,

$$\left\lceil \frac{n}{2} \right\rceil - e_i - 1 > n - 1 - e_i - (c+1)(n-c-1) \cdot \frac{2}{n+1}$$
$$2c^2 + 2(2-n)c + n^2 - n + 2 - \left\lceil \frac{n}{2} \right\rceil \cdot (n+1) < 0.$$

The discriminant of the left side with respect to c is equal to $4\left(-n^2 - 2n + 2\lceil \frac{n}{2} \rceil \cdot (n+1)\right)$. When n is even, $\lceil \frac{n}{2} \rceil = \frac{n}{2}$ and this expression equals $-4n$; therefore, the inequality above does not have a solution. If n is odd, $\lceil \frac{n}{2} \rceil = \frac{n+1}{2}$ and the inequality has the solution $\frac{n-3}{2} < c < \frac{n-1}{2}$. However, this interval does not contain integer numbers. In conclusion, such C cannot exist when $f' < \lceil \frac{n}{2} \rceil - e_i$. □

We are now ready to show $\mathcal{M}_1$ is NOM.

Proof of Theorem 1. We prove both Condition 1 and 2 of the definition of NOM hold true for any agent i.

Condition 1. This condition is the simplest to show. By Lemma 6, if i truthfully reports, the best outcome is attained by a partition where i is in a coalition with all her friends and no enemies. This coalition provides her the highest possible utility, so, Condition 1 holds true as no misreport can guarantee a strictly higher utility.

Condition 2. To understand the worst-case scenario we will make a case distinction depending on the number of enemies i has.

If $e_i \geq \lceil \frac{n}{2} \rceil$, for any $d_i \in \mathcal{D}_i$, consider a declaration of the others $\mathbf{d}_{-i}$ such that the friendship graph G^f for the resulting instance $(d_i, \mathbf{d}_{-i})$ is an i-centered octopus graph with the head $H = E_i$. By Lemma 3, for this instance, $\mathcal{M}_1$ outputs a partition containing $E_i \cup \{i\}$. This is the worst possible coalition for i according to her truthful declaration t_i. Since for any possible reporting of i there always exists a declaration of the others such that i ens up in a coalition together with all her enemies, Condition 2 is satisfied.

Assume now $e_i < \lceil \frac{n}{2} \rceil$. Let us choose $X \subseteq F_i$ such that $|E_i \cup X| = \lceil \frac{n}{2} \rceil$. We first show that, regardless of the declaration of i, there always exists an instance where the unique optimum has $E_i \cup X \cup \{i\}$ as a coalition. Given $d_i \in \mathcal{D}_i$, let us choose $\mathbf{d}_{-i}$ in such a way that $G^f(d_i, \mathbf{d}_{-i})$ is an i-centered octopus graph with the head $H = E_i \cup X$. Being $|H| \geq \lceil \frac{n}{2} \rceil$, by Lemma 3 we have that, in the unique optimum for this instance, $H \cup \{i\}$ is a coalition and remaining agents form singleton coalitions. By Lemma 6, such an outcome is the worst possible when declaring t_i, therefore, misreporting cannot improve the worst-case. This concludes our proof. □

4 COMPUTING THE OPTIMUM IS NP-HARD

Despite the existence of an optimum and NOM mechanism, in this section we show that computing an optimum partition is NP-hard.

Theorem 2. *For* FA *preferences, computing the optimum is NP-hard.*

We prove Theorem 2 with a reduction from the 3-Partition problem, which can be formulated as follows:

> **3-Partition problem**
>
> *Input:* A ground set $\{x_1, x_2, \ldots, x_{3m}\}$ of $3m$ elements such that
> (i) $\sum_{h=1}^{3m} x_h = mT$, for some $T > 0$;
> (ii) $x_h \in \mathbb{N}$, for each $h \in [3m]$;
> (iii) $\frac{T}{4} < x_h < \frac{T}{2}$, for each $h \in [3m]$.
>
> *Question:* Does there exist a partition of the ground set into m disjoint subsets $S_1, \ldots, S_m$ such that, for every $k \in [m]$, $S_k = \{s_k^1, s_k^2, s_k^3\}$ and $s_k^1 + s_k^2 + s_k^3 = T$?

Let us note that in the standard formulation of 3-Partition, condition (iii) is usually not required, however, the problem remains strongly NP-hard even under such a condition [22]. Moreover, condition (iii) also implies that for any $S \subseteq \{x_1, x_2, \ldots, x_{3m}\}$ if $\sum_{x \in S} x = T$, then $|S| = 3$. Consequently, any partition into subsets, each having sum T, is a partition into triples.

Reduction. Given a 3-Partition instance, let us construct the friendship graph G^f representing the corresponding FA instance.
Element-cliques: Each of these cliques represents a specific element in the ground set of the 3-Partition instance. In particular, for every $h \in [3m]$, we create a bidirectional clique K^h of size x_h.
Set-cliques: We create m bidirectional cliques $K_X^1, \ldots, K_X^m$ each one being of size $X = 4m^2T$. The choice of X is made in such a way that we can use the cliques $K_X^1, \ldots, K_X^m$ to interpret a coalition in an optimum partition, for the FA instance, as a set in the partition of the ground set for the 3-Partition instance.
Connections between cliques: We add x_h bidirectional edges between K^h and each K_X^k in such a way that there is exactly one bidirectional edge between each vertex of K^h and some node in K_X^k. Since $|K^h| = x_h < X$, this is always possible.

Notice that the number of agents is $n = \sum_{h=1}^{3m} x_h + mX = mT + 4m^3T$; thus, with 3-Partition being strongly NP-hard the correctness of the reduction proves the NP-hardness of our problem.

The optimum in the reduced instance. As a first step to prove Theorem 2, we need to understand the structure of a socially optimum outcome for the reduced instance. Thanks to a number of auxiliary lemmas, deferred to the full paper, we can conclude the following:

Proposition 1. *In the reduced instance, any socially optimum partition π^* is made by exactly m coalitions and, for each $C \in \pi^*$,*

(a) there exists a unique $k \in [m]$ such that $K_X^k \subseteq C$, and
(b) for every $h \in [3m]$, either $K^h \subseteq C$ or $K^h \cap C = \emptyset$.

To give an intuition, we chose X sufficiently large so that putting two cliques of size X in the same coalition would introduce too many negative relationships. Moreover, the positive connections between a clique (of both types, K_X^k or K^h) are particularly sparse if compared to their size, which guarantees that these cliques will not be split in any optimum outcome.

Proposition 1 turns out to be very helpful in restricting the outcomes that may possibly be optimum: in an optimum outcome, there are exactly m coalitions $C_1, \ldots, C_m$ each one consisting of a clique K_X^k and possibly some of the cliques $\{K^1, \ldots, K^{3m}\}$. Let us denote by Σ the set of such partitions and then let us make some important observations about any $\pi \in \Sigma$, which will help us to establish the optimum social welfare. Let us count the number of friendship and enemy relationships within coalitions of π:

(1) The cliques never get split, and hence, inside the coalitions of π there are always exactly $\alpha = mX(X-1) + \sum_{h=1}^{3m} x_h(x_h - 1)$ friendship relations between the members of the cliques;
(2) any element-clique is in a coalition with exactly one set-clique K_X^k, thus, the total number of the friendship relations between element- and set-cliques within the coalitions of π is constant and equals to $\beta = \sum_{h=1}^{3m} 2x_h = 2mT$;
(3) similarly, the total number of enemy relations between these groups is also always equaling $\gamma = \sum_{i=1}^{3m} 2x_h \cdot (X - x_h)$.

It remains to determine the enemy relationships between element-cliques that are in the same coalition. Assume w.l.o.g. that $\pi = (C_1, \ldots, C_m)$ and $K_X^k \subseteq C_k$ and denote by $s_k = \sum_{h:\, K^h \subseteq C_k} x_h$. Hence, the enemy relationships between element-cliques are $\sum_{k=1}^{m} s_k(s_k - 1) - \sum_{h=1}^{3m} x_h(x_h - 1)$. Putting all together, $\forall \pi \in \Sigma$,

$$\mathrm{SW}(\pi) = \alpha + \beta - \frac{\gamma}{n} - \frac{1}{n}\left(\sum_{k=1}^{m} s_k(s_k - 1) - \sum_{h=1}^{3m} x_h(x_h - 1)\right).$$

The only aspect affecting the SW of $\pi \in \Sigma$ is how $K^1, \ldots, K^{3m}$ are located in its coalitions. In particular, the social welfare is maximum when $\sum_{k=1}^{m} s_k(s_k - 1)$ is minimum. Such a quantity is minimized for $s_1 = s_2 = \cdots = s_m = T$. In conclusion, the 3-Partition instance is a "yes" instance if and only if in the social optimum of the reduced FA we have $s_1 = \cdots = s_m = T$. This proves Theorem 2.

5 AN APPROXIMATION MECHANISM

For the sake of achieving NOM in polynomial time, in this section, we present a $(4 + o(1))$-approximation mechanism. We recall that in [19] it was shown that creating a coalition for each weakly connected component of G^f is SP and guarantees an n-approximation to the optimum. This is so far the best approximation achieved by an SP mechanism. The bad performances of this mechanism can be attributed to the fact that when G^f is weakly connected but really sparse it would be convenient to split the unique weakly connected component of G^f into smaller coalitions, see for instance Example 3. To circumvent this problem, in [19], the authors presented a randomized mechanism, which we will hereafter call RandMech. It randomly splits the agents into two sets, each agent having probability $\frac{1}{2}$ of being in one of them, and then computes the weakly connected components on the two sides. RandMech is SP (in expectation) and guarantees an expected approximation ≤ 4.

Inspired by RandMech, we draw our deterministic and NOM mechanism. Specifically, it partitions the agents into two sets, P_1 and P_2, of size $\lceil \frac{n}{2} \rceil$ and $\lfloor \frac{n}{2} \rfloor$, respectively. It then updates P_1 and P_2, through the subroutine ImproveSW more formally described in the full paper. ImproveSW repeatedly tries to improve $\mathrm{SW}(\{P_1, P_2\})$ by swapping two agents, that is, simultaneously moving $i \in P_1$ to P_2 and $j \in P_2$ to P_1, or moving an agent from the largest to the smallest coalition (in case the two sets have the same size the algorithm will never perform move). ImproveSW terminates when no swap or move can increase the SW. The mechanism then

computes the weakly connected components in P_1 and P_2 which will be the coalitions of the returned coalition structure.

To show the mechanism is NOM, the initialization of $\{P_1, P_2\}$ will be crucial. Recall that $\delta(i)$ is the number of nodes weakly connected to i, while $N(X)$, for $X \subseteq \mathcal{N}$, is the set of agents weakly connected to X. The mechanism will create the initial $\{P_1, P_2\}$ by greedily adding agents to the set P_1 in the following way: At first, it inserts an agent $i \in \mathcal{N}$ with highest $\delta(i)$, then, iteratively proceeds by including an agent $j \in N(P_1) \setminus P_1$ with highest $\delta(j)$ – ties broken arbitrarily. This process continues until P_1 contains exactly $\lceil \frac{n}{2} \rceil$ agents. If at some point $N(P_1) \setminus P_1 = \emptyset$, the mechanism selects a new agent $i \in \mathcal{N} \setminus P_1$ with highest $\delta(i)$, and proceeds as before. We call this partition a *greedy 2-partition* of $\mathcal{N}$. In summary:

Mechanism $\mathcal{M}_2$. Given a set of agents $\mathcal{N}$ and their declarations $\mathbf{d}$, the mechanism creates a greedy 2-partition $\{P_1, P_2\}$. Then, while possible, it updates the partition using ImproveSW: $\{P_1, P_2\} \leftarrow$ ImproveSW(P_1, P_2). Finally, it computes $C_1, \ldots, C_m$, the weakly connected components of P_1 and P_2, and returns $\pi = \{C_1, \ldots, C_m\}$.

Theorem 3. *For* FA *instances, Mechanism $\mathcal{M}_2$ is NOM and guarantees a $(4 + o(1))$-approximation of the optimum in polynomial time.*

Let $\pi^{\mathcal{M}_2} = \{C_1, \ldots, C_m\}$ be the outcome of $\mathcal{M}_2$. We denote by f_π the number of friendships within the coalitions in a partition π.

Observation 1. *If π and π' are the partitions before and after the execution of a swap or move step by ImproveSW during $\mathcal{M}_2$, then,*

$$\mathrm{SW}(\pi') - \mathrm{SW}(\pi) = (f_{\pi'} - f_\pi)\,(1 + 1/n)\ .$$

Proof. Let $s = \lceil \frac{n}{2} \rceil \cdot \left(\lceil \frac{n}{2} \rceil - 1\right) + \lfloor \frac{n}{2} \rfloor \cdot \left(\lfloor \frac{n}{2} \rfloor - 1\right)$ be the total number of possible connections within the coalitions of π. We notice that s is also the number of possible connections within the coalitions of π'. In fact, a swap or move executed by the mechanism does not change the sizes and the number of coalitions. Therefore, $\mathrm{SW}(\pi) = f_\pi \left(1 + \frac{1}{n}\right) - \frac{s}{n}$, and the same holds for $\mathrm{SW}(\pi')$ replacing f_π with $f_{\pi'}$. Hence, the thesis follows. □

In other words, a swap or a move in $\{P_1, P_2\}$ is convenient for the social welfare if and only if it strictly increases the number of positive relationships within coalitions. Such an observation implies that at most $f \leq n(n-1)$ swaps and moves will occur, therefore, the mechanism is polynomial. We next show $\mathcal{M}_2$ is NOM.

Proof $\mathcal{M}_2$ is NOM. We show that for any agent i there is no incentive to misreport her true preferences to improve either the best- or the worst-case scenario showing that both Condition 1 and 2 of the definition of NOM hold true.

Condition 1. Recall that F_i is the truthful set of i's friends of size f_i. We start by noticing that, regardless of d_i, in any outcome of the mechanism i cannot be put in a coalition with more than $\min\{f_i, \lceil \frac{n}{2} \rceil - 1\}$ friends; in fact, no coalition can have more than $\lceil \frac{n}{2} \rceil$ agents. We next show that if i truthfully reports t_i, there exists $\mathbf{d}_{-i}$ such that i gets in $\mathcal{M}_2(t_i, \mathbf{d}_{-i})$ utility equal to $\min\{f_i, \lceil \frac{n}{2} \rceil - 1\}$.

If $f_i \leq \lceil \frac{n}{2} \rceil - 1$, let $\mathbf{d}_{-i}$ be such that F_i is a bidirectional clique in $G^f(t_i, \mathbf{d}_{-i})$ and all agents in F_i consider i a friend. The remaining agents are isolated. In this case, it is easy to see, the mechanism will output $F_i \cup \{i\}$ in a coalition and the other agents in singletons.

If $f_i > \lceil \frac{n}{2} \rceil - 1$, let $A \subseteq F_i$ such that $|A| = \lceil \frac{n}{2} \rceil - 1$. We choose $\mathbf{d}_{-i}$ so that A is a bidirectional clique and i is a friend for j if and only if $j \in A$. No other friendship relationship exists. In this case, when initializing P_1, P_2, the mechanism will set $P_1 = A \cup \{i\}$ and $P_2 = \mathcal{N} \setminus P_1$. In fact, for any agent $j \notin A \cup \{i\}$, we have $\delta(j) \leq 1$ while $A \cup \{i\}$ is weakly connected and each agent $j' \in A \cup \{i\}$ has $\delta(j') = |A| > 1$. The subroutine ImproveSW will not change the partition $\{P_1, P_2\}$ as no agent in A is weakly connected to P_2 and the number of positive relations between i and A is higher than for the ones between i and P_2, hence, no swap or move is profitable for the SW.

Putting all together, by truthfully reporting, it is possible for i to end up in a coalition of value $\min\{f_i, \lceil \frac{n}{2} \rceil - 1\}$. Since, regardless of the declaration of i, she cannot achieve a strictly higher utility in an outcome of $\mathcal{M}_2$, Condition 1 of NOM is satisfied.

Condition 2. Recall e_i is the size of the truthful set of enemies E_i.

If $e_i \geq \lceil \frac{n}{2} \rceil - 1$, given $A \subseteq E_i$ with $|A| = \lceil \frac{n}{2} \rceil - 1$, for any possible declaration d_i of i, we select $\mathbf{d}_{-i}$ so that A is a bidirectional clique and i is a friend for j if and only if $j \in A$. No other friendship relationship exists. Also in this case, the mechanism will set $P_1 = A \cup \{i\}$ and $P_2 = \mathcal{N} \setminus P_1$ and no swap and move will take place. Being P_1 weakly connected, i will end up in a coalition with $\lceil \frac{n}{2} \rceil - 1$ enemies. Since the mechanism outputs coalitions of size at most $\lceil \frac{n}{2} \rceil$, in no outcome of $\mathcal{M}_2$ i has a worse utility. Therefore, no misreport of i can guarantee her a strictly better worst-case.

If $e_i < \lceil \frac{n}{2} \rceil - 1$, which implies, $e_i \leq \lfloor \frac{n}{2} \rfloor - 1$, we start by showing that if i truthfully reports t_i, then, i cannot be in a coalition with less than $\lfloor \frac{n}{2} \rfloor - e_i - 1$ friends. In fact, let $\pi = \{P_1, P_2\}$ be the partition before $\mathcal{M}_2$ computes the weakly connected components. Assume $i \in P_h$, for some $h \in [2]$. Then, the number of agents other than i in P_h is at least $\lfloor \frac{n}{2} \rfloor - 1$. Thus, $|F_i \cap P_h|$ is minimized when $|P_h \setminus \{F_i \cup \{i\}\}| = e_i$, and hence i has at least $\lfloor \frac{n}{2} \rfloor - e_i - 1$ friends in P_h. Since the coalition of i will be the weakly connected component of i in P_h, i will be put in a coalition containing $F_i \cap P_h$. In conclusion, whatever the partition of i is, by truthfully reporting, i will always be in a coalition with at least $\lfloor \frac{n}{2} \rfloor - e_i - 1$ many friends.

We next show that, regardless of the declaration of i, there exists $\mathbf{d}_{-i}$ such that i is put in a coalition with all her enemies and $\lfloor \frac{n}{2} \rfloor - e_i - 1$ many friends. Let $A = \mathcal{N} \setminus \{E_i \cup X \cup \{i\}\}$ where $X \subset F_i$ such that $|E_i \cup X \cup \{i\}| = \lfloor \frac{n}{2} \rfloor$, which implies, $|A| = \lceil \frac{n}{2} \rceil$. Let $\mathbf{d}_{-i}$ be such that A is a bidirectional clique, all agents in E_i consider i a friend. In this case, initially, the mechanism will set $P_1 = A \setminus \{j\} \cup \{i\}$, for some $j \in A$. In fact, for this instance, $\delta(i) = n - 1$, $\delta(j') = |A| > 1$, for all $j' \in A$, and $\delta(j'') = 1$, for all $j'' \in \mathcal{N} \setminus \{A \cup \{i\}\}$. Moreover, $A \cup \{i\}$ is weakly-connected, hence, computing the greedy 2-partition the mechanism selects at first i and then $\lceil \frac{n}{2} \rceil - 1$ agents in A. The mechanism will then improve the social welfare of $\{P_1, P_2\}$ with ImproveSW. ImproveSW(P_1, P_2) will perform only the swap of i and j as this strictly increases the number of friendship relationships within coalitions – no agent in A considers i a friend while all of them consider j a friend, and j is not connected to any agent in $A \cup \{i\}$ while i possibly is. Once $P_1 = A$, no other swap or move will occur – a swap or move cannot strictly increase the number of positive relationships within coalitions. Therefore, the mechanisms will compute the weakly connected components in P_1 and P_2. Since $P_2 = E_i \cup X \cup \{i\}$ is weakly connected, i will be put in

the coalition P_2, and this is so, regardless of the declaration of i. This scenario is the worst possible for i among the possible outcomes of $\mathcal{M}_2$: in fact, we have shown that i cannot have strictly fewer friends and in this coalition the number of enemies is maximum. In conclusion, there is no way for i to increase the worst outcome by misreporting her preferences. This shows that Condition 2 is satisfied and concludes our proof. □

To determine the approximation ratio of $\mathcal{M}_2$, we need to establish a lower bound for $f_{\pi^{\mathcal{M}_2}}$, the number of friendships within the coalitions of $\pi^{\mathcal{M}_2}$, w.r.t. the overall friendship relationships f.

Lemma 7. *For an* FA *instance with f friendships, $f_{\pi^{\mathcal{M}_2}} \geq \frac{n-2}{2n-1} f$.*

Proof. Let f_{S_1,S_2} be the number of edges between S_1 and S_2 and f_S be the number of edges within S, for $S_1, S_2, S \subseteq \mathcal{N}$.

Let $\{P_1, P_2\}$ be the output of ImproveSW during the execution of $\mathcal{M}_2$. When splitting $\{P_1, P_2\}$ into weakly connected components the number of friendship relationships within coalitions remains the same. Therefore, $f_{\pi^{\mathcal{M}_2}}$ equals $f_{P_1} + f_{P_2}$. Moreover, when ImproveSW terminates, a swap of two agents of P_1 and P_2 does not increase the SW. Thus, for every $i \in P_1$ and $j \in P_2$,

$$f_{\{i\},P_2\setminus\{j\}} - f_{\{i\},P_1\setminus\{i\}} + f_{\{j\},P_1\setminus\{i\}} - f_{\{j\},P_2\setminus\{j\}} \leq 0\,,$$

as, from Observation 1, a swap is performed as long as it strictly increases the number of friendships within P_1 and P_2.

Summing up these inequalities for all $i \in P_1$ and $j \in P_2$:

$$(|P_2|-1)\cdot f_{P_1,P_2} - 2|P_2|\cdot f_{P_1} + (|P_1|-1)\cdot f_{P_2,P_1} - 2|P_1|\cdot f_{P_2} \leq 0\,.$$

Since $f_{P_1,P_2} = f - f_{\pi^{\mathcal{M}_2}}$, $|P_1| = \lceil \frac{n}{2} \rceil$, and $|P_2| = \lfloor \frac{n}{2} \rfloor$,

$$\left(\left\lceil \frac{n}{2} \right\rceil + \left\lfloor \frac{n}{2} \right\rfloor - 2\right)\cdot (f - f_{\pi^{\mathcal{M}_2}}) - 2\left\lfloor \frac{n}{2} \right\rfloor \cdot f_{P_1} - 2\left\lceil \frac{n}{2} \right\rceil \cdot f_{P_2} \leq 0$$

and, using $\lfloor \frac{n}{2} \rfloor \leq \lceil \frac{n}{2} \rceil \leq \frac{n+1}{2}$ and $\lfloor \frac{n}{2} \rfloor + \lceil \frac{n}{2} \rceil = n$, we finally derive

$$(n-2)(f - f_{\pi^{\mathcal{M}_2}}) \leq 2\left\lceil \frac{n}{2} \right\rceil \left(f_{P_1} + f_{P_2}\right) \leq (n+1)\cdot f_{\pi^{\mathcal{M}_2}}\,.$$

In conclusion, $\frac{2n-1}{n-2} f_{\pi^{\mathcal{M}_2}} \geq f$. □

This lemma constitutes the bulk of the proof of the approximation guarantee. We defer the details to the full paper. In a nutshell, we use essentially the same analysis as was made for RandMech, which has an expected approximation ratio of at most 4. RandMech, however, guarantees that in expectation exactly $f/2$ positive edges are within coalitions of the outcome, while, by the above lemma, we only have that at least $f \cdot \left(\frac{1}{2} - o(1)\right)$ positive relations are within coalitions, which leads to the approximation factor of $4 + o(1)$. In the full paper, we also show that there exists an instance where the approximation factor of $\mathcal{M}_2$ is $4 - o(1)$.

6 ENEMIES AVERSION PREFERENCES

Enemies Aversion (EA) preferences are the counterpart of FA where agents give priority to coalitions with fewer enemies, and when the number of enemies is the same, they prefer coalitions with a higher number of friends. This can be encoded in the class of ASHG with values $v_i(j) = \frac{1}{n}$, if $j \in F_i$, and $v_i(j) = -1$, otherwise. For this class, in [19], it has been shown that the optimum is hard to approximate within a factor $O(n^{1-\epsilon})$, for any positive ϵ. Moreover, a poly-time and $O(n)$-approximating SP mechanism exists but, even if the time complexity is not a concern, strategyproofness and optimality are not compatible. It is therefore natural to wonder if optimality is compatible with NOM. Unfortunately, this is not the case.

Theorem 4. *For* EA *preferences, no optimum mechanism is NOM*

Proof. Let us consider an instance where agent i has exactly one enemy in E_i, the truthful set of enemies. If all other agents declare everyone else as friends, the grand coalition is the social optimum and i gets utility of $\frac{n-2}{n} - 1$. We do not know if this is the worst case, but this means that at least one outcome where when reporting truthfully i ends up with strictly negative utility exists.

Assume now agent i declares everyone is her enemy. Then, in any social optimum, i is put in the singleton coalition and obtains utility 0. So, this manipulation improves the worst case and violates Condition 2 in the definition of NOM, which makes the mechanism outputting the social optimum obviously manipulable. □

It has also been proven that an SP and $(1+\sqrt{2})$-approximating mechanism exists [19]. It would be interesting to further investigate what are the boundaries of approximating SP or NOM mechanisms when time complexity is out of discussion.

7 CONCLUSIONS

In this paper, we investigated NOM in HGs with FA preferences, aiming at designing mechanisms optimizing the social welfare while preventing manipulation. Despite proving that computing a welfare-maximizing partition is NP-hard, we showed that a NOM mechanism computing the optimum always exits. In turn, for EA preferences, such a mechanism cannot exist. To address the computational challenges of optimal outcomes under FA, we presented a $(4+o(1))$-approximation mechanism that is NOM and runs in polynomial time. This mechanism not only improves on the best-known strategyproof mechanism, which provides a linear approximation in the number of agents, but also represents the first deterministic constant-factor approximation algorithm for FA preferences; this is an interesting contrast to EA preferences for which it is hard to approximate the optimum within a factor of $O(n^{1-\epsilon})$.

Interesting future research directions include the study of NOM for more general classes of HGs: for example, in ASHGs no bounded approximation is possible when requiring SP, so, it would be natural to consider a weaker notion of manipulability. Conversely, future work may focus on desirable properties like stability, welfare maximization (even beyond the utilitarian welfare), efficiency, or fairness, determining which kind of manipulations they are sensitive to.

ACKNOWLEDGMENTS

This work was partially supported by: PNRR MIUR project FAIR - Future AI Research (PE00000013), Spoke 9 - Green-aware AI; PNRR MIUR project VITALITY (ECS00000041), Spoke 2 - Advanced Space Technologies and Research Alliance (ASTRA); the European Union - Next Generation EU under the Italian PNRR, Mission 4, Component 2, Investment 1.3, CUP J33C22002880001, partnership on "Telecommunications of the Future"(PE00000001 - program "RESTART"), project MoVeOver/SCHEDULE ("Smart interseCtions witH connEcteD and aUtonomous vehicLEs", CUP J33C22002880001); and GNCS-INdAM.

REFERENCES

[1] Georgios Amanatidis, Georgios Birmpas, George Christodoulou, and Evangelos Markakis. 2017. Truthful allocation mechanisms without payments: Characterization and implications on fairness. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. 545–562.

[2] Haris Aziz, Felix Brandt, and Paul Harrenstein. 2013. Pareto optimality in coalition formation. *Games and Economic Behavior* 82 (2013), 562–581.

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

[4] Haris Aziz and Alexander Lam. 2021. Obvious Manipulability of Voting Rules. In *Algorithmic Decision Theory - 7th International Conference, ADT 2021, Toulouse, France, November 3-5, 2021, Proceedings (Lecture Notes in Computer Science, Vol. 13023)*. Springer, 179–193.

[5] Suryapratim Banerjee, Hideo Konishi, and Tayfun Sönmez. 2001. Core in a simple coalition formation game. *Social Choice and Welfare* 18, 1 (2001), 135–153.

[6] Nathanaël Barrot, Kazunori Ota, Yuko Sakurai, and Makoto Yokoo. 2019. Unknown agents in friends oriented hedonic games: Stability and complexity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1756–1763.

[7] Francis Bloch and Effrosyni Diamantoudi. 2011. Noncooperative formation of coalitions in hedonic games. *International Journal of Game Theory* 40, 2 (2011), 263–280.

[8] Anna Bogomolnaia and Matthew O. Jackson. 2002. The Stability of Hedonic Coalition Structures. *Games and Economic Behavior* 38, 2 (2002), 201–230.

[9] Florian Brandl, Felix Brandt, Manuel Eberl, and Christian Geist. 2018. Proving the incompatibility of efficiency and strategyproofness via SMT solving. *Journal of the ACM (JACM)* 65, 2 (2018), 1–28.

[10] Jiehua Chen, Gergely Csáji, Sanjukta Roy, and Sofia Simola. 2023. Hedonic Games With Friends, Enemies, and Neutrals: Resolving Open Questions and Fine-Grained Complexity. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*. ACM, 251–259.

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

[12] Dinko Dimitrov and Shao Chin Sung. 2004. Enemies and Friends in Hedonic Games: Individual Deviations, Stability and Manipulation. *SSRN Electronic Journal* (02 2004). https://doi.org/10.2139/ssrn.639483

[13] Jacques H Dreze and Joseph Greenberg. 1980. Hedonic coalitions: Optimality and stability. *Econometrica: Journal of the Econometric Society* (1980), 987–1003.

[14] Edith Elkind, Angelo Fanelli, and Michele Flammini. 2020. Price of pareto optimality in hedonic games. *Artificial Intelligence* 288 (2020), 103357.

[15] Edith Elkind and Michael J Wooldridge. 2009. Hedonic coalition nets.. In *AAMAS (1)*. Citeseer, 417–424.

[16] Moran Feldman, Liane Lewin-Eytan, and Joseph (Seffi) Naor. 2015. Hedonic Clustering Games. *ACM Trans. Parallel Comput.* 2, 1 (may 2015), 4:1–4:48.

[17] Michele Flammini, Bojana Kodric, Gianpiero Monaco, and Qiang Zhang. 2021. Strategyproof mechanisms for additively separable and fractional hedonic games. *Journal of Artificial Intelligence Research* 70 (2021), 1253–1279.

[18] Michele Flammini, Bojana Kodric, Martin Olsen, and Giovanna Varricchio. 2021. 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*, Vol. 12607. Springer, 159–174.

[19] Michele Flammini, Bojana Kodric, and Giovanna Varricchio. 2022. Strategyproof mechanisms for Friends and Enemies Games. *Artif. Intell.* 302 (2022), 103610.

[20] Michele Flammini and Giovanna Varricchio. 2022. Approximate Strategyproof Mechanisms for the Additively Separable Group Activity Selection Problem. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*. 300–306.

[21] Martin Gairing and Rahul Savani. 2010. Computing stable outcomes in hedonic games. In *International Symposium on Algorithmic Game Theory*. Springer, 174–185.

[22] Michael R Garey and David S Johnson. 1979. *Computers and intractability*. Vol. 174. freeman San Francisco.

[23] Jana Hajduková. 2004. On coalition formation games. *IM Preprints series A* 5 (2004).

[24] Ayumi Igarashi and Edith Elkind. 2016. Hedonic games with graph-restricted communication. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 242–250.

[25] Anna Maria Kerkmann, Nhan-Tam Nguyen, Anja Rey, Lisa Rey, Jörg Rothe, Lena Schend, and Alessandra Wiechers. 2022. Altruistic hedonic games. *Journal of Artificial Intelligence Research* 75 (2022), 129–169.

[26] Bettina-Elisabeth Klaus, Flip Klijn, and Seçkin Özbilen. 2023. Core Stability and Strategy-Proofness in Hedonic Coalition Formation Problems with Friend-Oriented Preferences. *Available at SSRN 4677609* (2023).

[27] Selçuk Özyurt and M Remzi Sanver. 2009. A general impossibility result on strategy-proof social choice hyperfunctions. *Games and economic behavior* 66, 2 (2009), 880–892.

[28] Alexandros Psomas and Paritosh Verma. 2022. Fair and Efficient Allocations Without Obvious Manipulations. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

[29] Carmelo Rodríguez-Álvarez. 2009. Strategy-proof coalition formation. *International Journal of Game Theory* 38 (2009), 431–452.

[30] Jörg Rothe, Hilmar Schadrack, and Lena Schend. 2018. Borda-induced hedonic games with friends, enemies, and neutral players. *Mathematical Social Sciences* 96 (2018), 21–36.

[31] Peter Troyan. 2024. (Non-)obvious manipulability of rank-minimizing mechanisms. *Journal of Mathematical Economics* (2024), 103015.

[32] Peter Troyan and Thayer Morrill. 2020. Obvious manipulations. *Journal of Economic Theory* 185 (2020), 104970.

[33] Giovanna Varricchio. 2023. On Approximate Strategyproof Mechanisms for Hedonic Games and the Group Activity Selection Problem. In *Proceedings of IPS (CEUR Workshop Proceedings, Vol. 3585)*. CEUR-WS.org.