Welfare Approximation in Additively Separable Hedonic Games

Martin Bullinger
University of Oxford
Oxford, United Kingdom
martin.bullinger@cs.ox.ac.uk

Vaggos Chatziafratis
University of California, Santa Cruz
Santa Cruz, United States
vaggos@ucsc.edu

Parnian Shahkar
University of California, Irvine
Irvine, United States
shahkarp@uci.edu

ABSTRACT

Partitioning a set of n items or agents while maximizing the value of the partition is a fundamental algorithmic task. We study this problem in the specific setting of maximizing social welfare in additively separable hedonic games. Unfortunately, this task faces strong computational boundaries: Extending previous results, we show that approximating welfare by a factor of $n^{1-\varepsilon}$ is NP-hard, even for severely restricted weights. However, we can obtain a randomized $\log n$-approximation on instances for which the sum of input valuations is nonnegative. Finally, we study two stochastic models of aversion-to-enemies games, where the weights are derived from Erdős-Rényi or multipartite graphs. We obtain constant-factor and logarithmic-factor approximations with high probability.

KEYWORDS

Algorithmic game theory; coalition formation; social welfare; approximation algorithms

ACM Reference Format:
Martin Bullinger, Vaggos Chatziafratis, and Parnian Shahkar. 2025. Welfare Approximation in Additively Separable Hedonic Games. 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

Partitioning a set of items or agents, say humans or machines, is a fundamental problem that has been studied across many disciplines such as computer science, economics, or mathematics. For instance, it is relevant in the context of clustering, an important task in machine learning with far-reaching applications like image segmentation [19], or for community detection, which helps in understanding networks, e.g., of societies or physical systems [31].

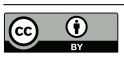
Our paper takes a game-theoretic perspective and considers the prominent model of *additively separable hedonic games* [6]. We assume that there is a set of agents that has to be partitioned into coalitions and agents have preferences over the coalitions that they are part of [21]. Preferences are given by a weighted graph, where the agents are the vertices and the edge weights encode the valuation between agents. The utility of an agent for a coalition is the sum of weights of edges towards members of this coalition. This class of games is quite expressive and contains more structured subclasses of games. For instance, an agent might divide the other agents into friends and enemies and could simply try to maximize the number of friends within their coalition while minimizing the number of enemies. A priority between these two objectives can be captured by the exact edge weights: for example, if there is a large negative weight for enemies and a small positive weight for friends, then minimizing enemies is much more important than maximizing friends, as conceptualized in so-called *aversion-to-enemies games* [20].

A fundamental quantity for evaluating a possible output is its *social welfare* (also called utilitarian welfare) which is the sum of all agents utilities. Unfortunately, maximizing this quantity faces significant computational boundaries. Aziz et al. [2] show that it is NP-hard to maximize, and, even worse, approximating maximum welfare by a factor of at least $n^{1-\varepsilon}$ is NP-hard for any $\varepsilon > 0$ [25]. Our paper aims at circumventing this computational boundary.

First, we investigate the inapproximability of maximum welfare. Notably, the result by Flammini et al. [25] is for aversion-to-enemies games, which use valuations $-n$ and 1, i.e., the negative valuation is dependent on the number of agents n. We complement this by showing an $n^{1-\varepsilon}$-inapproximability result on instances in which the valuations are restricted to $\{-v^-, 0, 1\}$, where $v^- \geq 1$ is an arbitrary but fixed (and, therefore, globally bounded) number.[1] This sounds discouraging but it strengthens the impression that negative valuations seem to be the reason for computational boundaries.

In the remainder of the paper, we provide several possibilities to achieve better approximation guarantees. First, we consider the restricted domain of games in which the sum of all valuations is nonnegative. This assumption still allows for the existence of rather negative valuations, however, it disallows an overall bias towards negative valuations. We make use of a result from the correlation clustering literature [17] to prove the existence of a randomized algorithm that approximates social welfare by a factor of $O(\log n)$.

Second, we consider two stochastic models of aversion-to-enemies games in which we achieve approximation guarantees with high probability. We start by assuming a basic model where valuations originate from an Erdős-Rényi graph. We show that a constant approximation of maximum welfare is possible. Subsequently, we define a stochastic model inspired by team management where every agent has a role, such as project manager, software engineer, UX designer, or marketing specialist. Coalitions represent teams and each role should be present in a team at most once. This scenario can be conceptualized by making agents with the same role mutually incompatible by introducing large negative valuations. In other words, the compatibility of agents is captured by a multipartite graph where the roles induce a partition of the vertices. However, in reality, even agents of different roles might be incompatible for various reasons. We model this by introducing a parameter p that captures the probability of agents being incompatible. In our stochastic model, every pair of agents admitting different



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).

[1] By rescaling valuations, this is equivalent to assuming that, in addition to a neutral valuation of 0, there is a single positive and negative valuation, where the former is bounded by the absolute value of the latter.

roles are incompatible independently. Based on the magnitude of p we obtain perturbation regimes that lead to different approximation guarantees. In the low perturbation regime, we can approximate maximum welfare by a constant factor, whereas a high perturbation regime allows for a $\log n$-approximation.

2 RELATED WORK

Hedonic games were introduced by Drèze and Greenberg [21] as an ordinal model of coalition formation, in which agents state their preferences as rankings over coalitions. Their broad consideration started, however, only 20 years later [4, 6, 15]. Much of their popularity today is due to the introduction of additively separable hedonic games by Bogomolnaia and Jackson [6] in this era.

While these first papers were in the realm of economic theory, they soon sparked a broader consideration of hedonic games in computer science. This led to increased attention of algorithmic properties of solution concepts, including their computational complexity [3, 14]. Social welfare was first realized to be a demanding objective by Aziz et al. [2] who showed that it is NP-hard to compute even if valuations are restricted to be only -1 or 1. Subsequently, Flammini et al. [25] significantly strengthened this to the $n^{1-\varepsilon}$-inapproximability result for aversion-to-enemies games mentioned in the introduction.

Beyond social welfare, other welfare objectives have been explored. Some early papers on hedonic games already studied Pareto optimality, a less demanding notion of welfare studied throughout economics [6, 21]. Pareto-optimal coalition structures can be computed in polynomial time under fairly general assumptions including symmetric valuations [2, 11]. However, this yields no approximation of social welfare because Pareto-optimal outcomes may have negative social welfare [22]. Moreover, Aziz et al. [2] also considered egalitarian welfare, which aims at maximizing the utility of the worst-off agent. Despite its challenges for the offline model, welfare approximation has also been studied in an online variant of additively separable hedonic games [13, 26]. Flammini et al. [26] consider a general model where no finite competitive ratio is possible if the utility range is unbounded. Moreover, Bullinger and Romen [13] study a model where the algorithm is allowed to dissolve coalitions into singleton coalitions, which allows to achieve a coalition structure with a social welfare that is at most a factor of $\Theta(n)$ worse than the maximum possible welfare. In particular, they show that maximum weight matchings achieve an n-approximation of social welfare [13]. This essentially matches the aforementioned inapproximability by a factor of $n^{1-\varepsilon}$ [25]. Finally, social welfare has been considered in a mechanism design perspective aiming at strategyproof preference elicitation [24, 25].

Beyond welfare, the most common objectives in hedonic games are notions of stability [2, 6–9, 27, 32, 34]. Rather than the global guarantees provided by welfare notions, stability assumes a more strategic perspective in that it requires the absence of beneficial deviations by single agents or groups of agents. Single-deviation stability often leads to NP-completeness [8, 32], whereas group stability can even be Σ_2^p-complete [34]. Interestingly, symmetric valuations lead to the existence of stable outcomes based on single-agent deviations [6], but their computation is still infeasible. It is PLS-complete, i.e., complete for the complexity class capturing problems that guarantee solutions based on local search algorithms [27]. An interesting objective that combines ideas of stability and global guarantees is popularity [2, 7], which is akin to weak Condorcet winners as studied in social choice theory [10].

While all the literature discussed so far considers a deterministic model, stochastic models have been studied to some extent [12, 23]. In particular, Bullinger and Kraiczy [12] show how to obtain stable outcomes if valuations are drawn uniformly at random. Their algorithm runs in three stages, the first of which will turn out to be useful in obtaining welfare guarantees as well, see Section 5.1. By contrast, Fioravanti et al. [23] consider a deterministic game model and aim at computing outcomes that are stable with high probability.

Finally, hedonic games are also related to other graph partitioning problems such as correlation clustering. The input typically consists of a complete graph with edges labeled as "+" or "−" to indicate similarity or dissimilarity, respectively [5, 18, 33]. The goal is to find a partition that maximizes agreements as measured by the sum of "+" edges inside clusters plus "−" edges across different clusters. Other objectives where the goal is to minimize errors of the partition, measured by "−" edges within clusters plus "+" edges across clusters have also been extensively studied [16]. By contrast, our social welfare objective in hedonic games is different in that it accounts only for the edges within the coalitions (in particular, it ignores the edges across different coalitions).

Going beyond worst-case analysis, the two stochastic models we study for hedonic games in the second part of our paper relate to various stochastic models with random (or semi-random) edges that have been proposed for correlation clustering. For example, [30] investigates a noisy model on complete graphs, where they start from an arbitrary partition of the vertices into clusters and for each pair of vertices, the edge information (either 1 or -1) is corrupted independently with probability p. Other average-case models and extensions to arbitrary graphs (not necessarily complete) have been studied by Makarychev et al. [29], with the goal of designing provably good approximation algorithms.

3 PRELIMINARIES

Consider a finite set N of $n := |N|$ agents. A *coalition* is a nonempty subset of N. We denote by $\mathcal{N}_i := \{S \subseteq N : i \in S\}$ the set of all coalitions that agent i belongs to. A *coalition structure* (or *partition*) is a partition π of N into coalitions, i.e., $\bigcup_{C\in\pi} C = N$ and for each pair of coalitions $C, C' \in \pi$ with $C \neq C'$ it holds that $C \cap C' = \emptyset$. Note that there is no bound on the number or size of coalitions. For an agent $i \in N$, we denote by $\pi(i)$ the coalition that i belongs to in π. We denote the set of all partitions of N by Π_N, and the set of all partitions containing exactly two coalitions as $\Pi_N^{(2)}$, i.e., $\Pi_N^{(2)} := \{\pi \in \Pi_N : |\pi| = 2\}$. A coalition is called a *singleton coalition* if it contains exactly one agent. The partition where every agent is in a singleton coalition is called the *singleton partition*.

In a hedonic game, every agent possesses preferences over the coalitions in $\mathcal{N}_i$. We use the model of additively separable hedonic games by Bogomolnaia and Jackson [6] in which these preferences are obtained from cardinal valuations that can be encoded by a complete and directed weighted graph. Formally, a *cardinal hedonic*

game is a pair (N, u) where $u = (u_i \colon \mathcal{N}_i \to \mathbb{R})_{i \in N}$ is a vector of *utility functions*. An *additively separable hedonic game* (ASHG) is specified by a vector $v = (v_i \colon N \to \mathbb{R})_{i \in N}$ of (single-agent) *valuation functions*. It is then defined as the cardinal hedonic game (N, u), where for agent $i \in N$ and coalition $C \in \mathcal{N}_i$, it holds that

$$u_i(C) := \sum_{j \in C} v_i(j).$$

In words, the utility of a coalition is derived from single-agent values, which are aggregated by summing the values of the agents in this coalition. Since valuation functions fully specify an ASHG, we also speak of the ASHG (N, v). Note that ASHGs can be encoded by a weighted graph where agents are vertices and edge weights are given by the valuations.

We extend utilities over coalitions to utilities over partitions by defining $u_i(\pi) := u_i(\pi(i))$. Given coalitions $C, C' \in \mathcal{N}_i$, we say that agent i *prefers* C over C' if $u_i(C) \geq u_i(C')$. Moreover, we say that i *strictly prefers* C over C' if $u_i(C) > u_i(C')$. We use the same terminology for partitions. Given an ASHG (N, v) and a partition π, we define its *social welfare* as

$$\mathcal{SW}(\pi) := \sum_{i \in N} u_i(\pi) = \sum_{C \in \pi \colon i, j \in C} v_i(j).$$

Hence, the social welfare is the sum of the utilities which, in an ASHG, is equivalent to the sum of all valuations between agents in the same coalition. We denote by π^* a partition that maximizes $\mathcal{SW}$.

ASHGs admit various interesting subclasses when restricting valuations. Following Dimitrov et al. [20], an ASHG (N, v) is called an *aversion-to-enemies game* if $v_i(j) \in \{-n, 1\}$ for all $i, j \in N$. An ASHG (N, v) is called *symmetric* if for each pair of agents $i, j \in N$, it holds that $v_i(j) = v_j(i)$. We write $v(i, j)$ for the symmetric valuation function between i and j. In this paper, we restrict attention to symmetric ASHGs.[2]

Consider an ASHG (N, v) and an approximation ratio $c \geq 1$. A partition π is said to provide a *c-approximation to maximum welfare* if $c \cdot \mathcal{SW}(\pi) \geq \mathcal{SW}(\pi^*)$. We denote by c-ApproxWelfare the computational problem of, given an ASHG, computing a partition with a c-approximation to maximum welfare.

We consider both deterministic and randomized algorithms and aim at efficient algorithms. For $c \geq 1$, a polynomial-time algorithm is called a *c-approximation algorithm* for maximizing welfare if it solves c-ApproxWelfare. For randomized algorithms, the expected running time has to be bounded by a polynomial and the produced partition has to provide a c-approximation to maximum welfare in expectation. Note that we allow (and frequently assume) that the factor c depends on n.

Finally, given an ASHG (N, v), we define its *total value* as

$$\mathcal{V}(N, v) := \sum_{i, j \in N} v_i(j).$$

We will obtain good approximation guarantees by restricting attention to ASHGs with nonnegative total value.

[2] In general ASHGs, symmetry is without loss of generality when reasoning about social welfare as the welfare remains the same if we replace $v_i(j)$ and $v_j(i)$ by $\frac{v_i(j)+v_j(i)}{2}$, see, e.g., [11]. However, this is not the case for aversion-to-enemies games as the symmetrization may leave this game class.

In this paper, we use $[k]$ to represent the set $\{1, \ldots, k\}$. Moreover, in asymptotic statements, we state logarithms without base. They can be assumed to have base e.

4 DETERMINISTIC GAMES

Recall that Aziz et al. [1] show that maximizing social welfare is NP-hard. Their result even holds for symmetric valuations restricted to $\{-1, 1\}$. Moreover, Flammini et al. [25] prove that approximating social welfare by a factor of $n^{1-\varepsilon}$ is NP-hard for aversion-to-enemies games, i.e., when valuations are in the set $\{-n, 1\}$. In this section, we will significantly deepen the understanding of welfare approximability around this result. First, we show that the result does not rely on unbounded negative weights by providing a reduction for valuations restricted to $\{-v^-, 0, 1\}$ where $v^- \geq 1$. In particular, this means that unbounded negative weights or negative weights of absolute value much larger than the value of positive weights are not necessary. Subsequently, we will show how to circumvent the inapproximability result for ASHG with nonnegative total value.

4.1 Welfare Inapproximability for Restricted Valuations

We now prove our inapproximability result.[3]

Theorem 4.1. *Let $\varepsilon > 0$ and $v^- \geq 1$. Then, unless* P = NP, *$n^{1-\varepsilon}$-ApproxWelfare cannot be solved in polynomial time for symmetric ASHGs with valuations in the set $\{-v^-, 0, 1\}$.*

Proof. Let $\varepsilon > 0$ and $v^- \geq 1$. We reduce from the $n^{1-\varepsilon}$-approximate MaximumClique problem. The input is an unweighted graph G and the task is to compute a clique C of G with $n^{1-\varepsilon} \cdot |C| \geq \mu^*$, where μ^* is the size of a maximum clique of G. Unless P = NP, this problem cannot be solved in polynomial time [35].

We now describe the reduction. Assume that we are given an unweighted graph $G = (V, E)$. We construct an ASHG (N, v) as follows. The set of players is $N = N_V \cup \{z\}$, where $N_V = \{a_u \colon u \in V\}$, i.e., N_V contains a player for each vertex of G. Symmetric valuations are given by

$$v(i, j) = \begin{cases} 1 & i = z, j \in N_V, \\ -v^- & i, j \in N_V, \{i, j\} \notin E, \text{ and} \\ 0 & i, j \in N_V, \{i, j\} \in E. \end{cases}$$

Let π be a partition of N and let $C_1 = \{u \in V \colon a_u \in \pi(z)\}$. Hence, C_1 is a vertex set in G. We create a sequence of vertex sets until we end with a clique of G. For $i \geq 1$, assume that we have constructed a set C_i. We stop if C_i is a clique. Otherwise, we find a vertex $u_i \in C_i$ such that u_i is not adjacent to all other vertices in C_i and set $C_{i+1} = C_i \setminus \{u_i\}$. Since the number of vertices of G is finite, this stops after $k \leq |V|$ steps with a clique C_k. For $1 \leq \ell \leq k$, we define the partition $\pi^\ell = \{\{a_u \colon u \in C_\ell\} \cup \{z\}\} \cup \{\{a_u\} \colon u \in V \setminus C_k\}$. We now show that, for $1 \leq \ell \leq k-1$, it holds that $\mathcal{SW}(\pi^\ell) \leq \mathcal{SW}(\pi^{\ell+1})$. Indeed, $v(u_\ell, u) = 1$ if $u = z$ and $v(u_\ell, u) \leq 0$ for all $u \in \pi^\ell(u_\ell) \setminus \{u_\ell, z\}$. Moreover, since C_ℓ is not a clique, there exists an agent $u \in \pi^\ell(u_\ell) \setminus \{u_\ell, z\}$ with $v(u_\ell, u) = -v^- \leq -1$. Hence, removing u_ℓ from their coalition and forming a singleton coalition can only increase the social welfare.

[3] We would like to thank Abheek Ghosh for the proof idea of Theorem 4.1.

In addition, it holds that $\mathcal{SW}(\pi) \leq \mathcal{SW}(\pi^1)$ since π^1 can only differ from π by dissolving nonsingleton coalitions not containing z, which can only increase the welfare. Finally, $\mathcal{SW}(\pi^k) = |C_k|$ as the only nonsingleton coalition in π^k is $\pi^k(z)$ which contains exactly $|C_k|$ other agents forming a clique in G. Hence,

$$\mathcal{SW}(\pi) \leq \mathcal{SW}(\pi^k) = |C_k|. \tag{1}$$

Next, let C^* be a maximum clique in G. Consider $\pi' = \{\{z\} \cup \{a_u : u \in C^*\}\} \cup \{\{a_u\} : u \in V \setminus C^*\}$ is a partition in (N, v) with $\mathcal{SW}(\pi') = |C^*|$. Hence, for the partition π^* maximizing welfare, it holds that

$$\mathcal{SW}(\pi^*) \geq \mathcal{SW}(\pi') = |C^*|. \tag{2}$$

Now assume that we have a polynomial-time algorithm computing a partition π with $n^{1-\varepsilon} \cdot \mathcal{SW}(\pi) \geq \mathcal{SW}(\pi^*)$. Clearly, the procedure described above to construct C_k runs in polynomial time as well. It holds that

$$n^{1-\varepsilon} \cdot |C_k| \overset{\text{Eq. (1)}}{\geq} n^{1-\varepsilon} \cdot \mathcal{SW}(\pi) \geq \mathcal{SW}(\pi^*) \overset{\text{Eq. (2)}}{\geq} |C^*|.$$

Hence, we have found a polynomial-time algorithm to approximate the maximum clique within a factor of $n^{1-\varepsilon}$. As argued in the beginning, this can only happen if P = NP, completing our proof. □

Notably, Theorem 4.1 immediately implies hardness of $n^{1-\varepsilon}$-ApproxWelfare for ASHGs with nonsymmetric valuations restricted to $\{-1, 1\}$. We can simply replace valuations $v(i, j) = 0$ by $v_i(j) = 1$ and $v_j(i) = -1$ and obtain a reduced instance in which all partitions have the identical welfare. However, it remains an open problem to resolve the complexity of c-ApproxWelfare for symmetric ASHGs with valuations restricted to $\{-1, 1\}$, even if $c > 1$ is assumed to be a constant not dependent on n.

4.2 Logarithmic Approximation for Nonnegative Total Value

We will now show that we can get beyond the inapproximability result of Theorem 4.1 if we restrict attention to ASHGs with a nonnegative total value. For this, we will draw a connection to a related problem from the literature on correlation clustering. Instances of correlation clustering usually only use binary information, i.e., whether two objects should belong to the same or different clusters. The goal is then to optimize one of two objectives: maximizing agreements, i.e., the number of pairs whose pairwise relationship is classified correctly, and minimizing disagreements, i.e., the number of pairs classified incorrectly. In addition, one can consider the combination of these two objectives, where agreements should be maximized while disagreements should simultaneously be minimized. In the spirit of hedonic games, we capture a weighted version of this objective as a notion of welfare. Given an ASHG (N, v) and a partition π, we define its *correlation welfare* as

$$\mathcal{CW}(\pi) := \frac{1}{2}\left[\sum_{i \in N}\left(\sum_{j \in \pi(i)} v_i(j) - \sum_{j \in N \setminus \pi(i)} v_i(j)\right)\right].$$

Charikar and Wirth [17] present a randomized $O(\log n)$-approximation algorithm for maximizing[4] $\mathcal{CW}$ subject to $\pi \in \Pi_N^{(2)}$.

[4] In their terminology, this is the problem of maximizing MaxQP.

They then show that this extends to maximizing $\mathcal{CW}$ within Π_N in the case of valuation functions in the range $\{-1, 1\}$. It is easy to see that the same technique applies for the general range of valuation functions (cf. Lemma 4.6). The goal of this section is to extend the approximation guarantee to maximizing $\mathcal{SW}$ for ASHGs with nonnegative total value.

By plugging in definitions, we immediately obtain the following relationship between $\mathcal{SW}$ and $\mathcal{CW}$. All proofs missing from this section can be found in the full version of our paper.

Proposition 4.2. *Consider an ASHG (N, v) and a partition π. Then it holds that $\mathcal{CW}(\pi) + \frac{1}{2}\mathcal{V}(N, v) = \mathcal{SW}(\pi)$.*

As a consequence, we obtain that the same partitions maximize $\mathcal{SW}$ and $\mathcal{CW}$.

Proposition 4.3. *Consider an ASHG (N, v). Then a partition maximizes $\mathcal{SW}$ if and only if it maximizes $\mathcal{CW}$.*

Hence, solving for social welfare maximization and correlation welfare maximization is exactly equivalent. However, this does not have any implications on approximation guarantees, as we illustrate in the next example.

Example 4.4. Let $x > 0$ be an arbitrary positive number. Consider the symmetric ASHG (N, v) with $N = \{a_1, a_2, a_3\}$ and symmetric valuation functions given by $v(a_1, a_2) = 1$, $v(a_1, a_3) = -x$, and $v(a_2, a_3) = 0$. Let π denote the singleton partition and $\pi^* = \{\{a_1, a_2\}, \{a_3\}\}$. Clearly, π^* is the unique partition maximizing $\mathcal{SW}$ and $\mathcal{CW}$. In addition, it holds that $\frac{1+x}{x-1}\mathcal{CW}(\pi) = \mathcal{CW}(\pi^*)$, whereas $\frac{\mathcal{SW}(\pi)}{\mathcal{SW}(\pi^*)} = 0$.

Now, consider any approximation guarantee $c > 1$. Then, π provides a c-approximation of $\mathcal{CW}$ for $x = \frac{c-1}{1+c}$, whereas π yields no c-approximation of $\mathcal{SW}$. ◁

The reason why approximate outcomes in terms of correlation welfare do not yield any guarantee on the social welfare in Example 4.4 is that there is a single valuation that is very negative. It is enough that the two corresponding agents are in different coalitions to obtain a good approximation of the correlation welfare. However, the picture changes if such a situation does not occur. If the total value of an instance is nonnegative, then social welfare inherits approximation guarantees from correlation welfare.

Lemma 4.5. *Let (N, v) be an ASHG such that $\mathcal{V}(N, v) \geq 0$. Let $c \geq 1$ and let π^* be a partition maximizing $\mathcal{CW}$. Let π be a partition with $c \cdot \mathcal{CW}(\pi) \geq \mathcal{CW}(\pi^*)$. Then it holds that $c \cdot \mathcal{SW}(\pi) \geq \mathcal{SW}(\pi^*)$.*

As a second lemma, we establish a relationship between maximizing welfare when partitions can only contain two coalitions and when partitions are unconstrained. Its proof is an adaptation of a similar result by Charikar and Wirth [17] concerning $\mathcal{CW}$ for valuations in $\{-1, 1\}$.

Lemma 4.6. *Let (N, v) be an ASHG with $\mathcal{V}(N, v) \geq 0$. Then it holds that $\max_{\pi \in \Pi_N} \mathcal{SW}(\pi) \leq 2 \cdot \max_{\pi \in \Pi_N^{(2)}} \mathcal{SW}(\pi)$.*

We can combine the two last lemmas to apply the main theorem by Charikar and Wirth [17] and obtain a randomized $O(\log n)$-approximation algorithm.

Theorem 4.7. *There exists a randomized $O(\log n)$-approximation algorithm for maximizing social welfare in ASHGs with nonnegative total value.*

Proof. Theorem 1 by Charikar and Wirth [17] states the existence of a randomized $O(\log n)$-approximation algorithm for $\mathcal{CW}$ under the constraint that partitions are in $\Pi_N^{(2)}$. By Lemma 4.5, the same approximation guarantee is obtained for $\mathcal{SW}$ under the same constraint. Finally, Lemma 4.6 guarantees that the maximum welfare of any partition is better by at most a factor of 2. □

5 BEYOND WORST-CASE ANALYSIS

In light of the hardness result by Flammini et al. [25] for approximating social welfare in aversion-to-enemies games, it is natural to ask how well we can approximate welfare in such games generated by stochastic models. In this section, we introduce two such models where the valuations originate from either Erdős-Rényi or multipartite graphs. Erdős-Rényi graphs serve as a common testbed for graph optimization problems and help us set the stage for the more challenging setting of multipartite graphs. Interestingly, our main theorems demonstrate that greedy algorithms are remarkably effective in these models, yielding constant-factor and logarithmic-factor approximations of social welfare.

In this section, $G = (N, v)$ refers to a fixed symmetric aversion-to-enemies game. In any partition of N, a valuation of $-n$ within a coalition implies a negative utility for the corresponding agents. Consequently, removing one of these agents from the coalition and forming a singleton coalition would increase the overall social welfare. This observation suggests that in an optimal partition π^*, no coalition contains agents with a mutual valuation of $-n$. Let G' denote the subgraph of G, obtained by removing all edges with weight $-n$. We now present a useful lemma.

Lemma 5.1. *If the size of the maximum clique in G' is t, then $\mathcal{SW}(\pi^*) \leq n(t-1)$.*

Proof. No coalition in the partition π^* contains an edge with weight $-n$. Therefore, each coalition in π^* forms a clique in G'. Since the size of a maximum clique in G' is t, the size of every coalition in π^* is at most t. Consequently, the utility of each agent is bounded by $t-1$. We conclude that

$$\mathcal{SW}(\pi^*) = \sum_{i \in N} u_i(\pi^*(i)) \leq n(t-1). \qquad \square$$

5.1 Erdős-Rényi Graphs

In our first model, we assume a set of agents, each pair of which is incompatible with probability $1-p$. We model this as a symmetric aversion-to-enemies game by assigning a valuation of $-n$ between incompatible agents and a valuation of 1 between compatible agents. This corresponds to sampling its underlying graph as follows.

Definition 5.2. A *weighted Erdős-Rényi graph* $G = (n, p)$ is a random weighted graph with n vertices such that, independently, each edge takes a weight of $-n$ with probability p and a weight of 1 with probability $1-p$.

We will show that a simple and natural greedy algorithm yields a constant-factor approximation of the maximum welfare with high probability. For this, we use the *greedy clique formation algorithm* by Bullinger and Kraiczy [12, Section 5.2] applied to the subgraph G' formed by removing all negative edges from a graph G. The algorithm greedily forms maximal cliques in G'. as long as the cliques reach a certain size threshold $t = \left\lceil \frac{\log_{1/p} n}{2} \right\rceil$. If, at any point, the size of the created maximal clique is smaller than t, the algorithm outputs the existing cliques as coalitions, and assigns singleton coalitions to the remaining agents. The following theorem measures the performance of this algorithm. It follows from the proof of Theorem 5.2 by Bullinger and Kraiczy [12].[5]

Theorem 5.3 (Bullinger and Kraiczy [12]). *Consider an Erdős-Rényi graph $G = (n, p)$ and let $b = \frac{1}{p}$. Then, with probability at least $1 - e^{-\Omega(\log_b^3 n)}$, the greedy clique formation algorithm assigns all except at most $\frac{n}{\log_b^2 n}$ to cliques of size $\left\lceil \frac{\log_b n}{2} \right\rceil$.*

We apply the theorem to obtain a constant-factor approximation of maximum welfare. Essentially, Theorem 5.3 allows to obtain a coalition with social welfare $\Theta(n \log n)$ while we apply Lemma 5.1 to show that the maximum welfare is also of this order.

Theorem 5.4. *Let $p \in (0, 1)$. Then there exists a constant-factor approximation algorithm for aversion-to-enemies games given by a weighted Erdős-Rényi graph $G = (n, p)$.*

Proof. Note that G' is a weighted Erdős-Rényi graph where every edge was sampled with probability p, i.e. $G' = (n, p)$. Let $b = \frac{1}{p}$ and let π be the resulting partition after applying the greedy clique formation algorithm to the subgraph G'. By Theorem 5.3, it follows that $\mathbb{E}[\mathcal{SW}(\pi)] = \Theta(n \log n)$.

In addition, the size of the maximum clique in G' is $O(\log_b n)$ with probability 1 [28]. Hence, by Lemma 5.1, the expected maximum welfare of a partition in G' is $O(n \log n)$. Note that all constants hidden in the asymptotic behavior only depend on b and, therefore, on p. It follows that the greedy clique formation algorithm yields a constant-factor approximation. □

To get a feeling on the constant hidden in the previous theorem, one can reason as follows. For large enough n, with probability $1 - \frac{1}{n}$, greedy clique formation will place $\frac{2n}{3}$ agents into coalitions of size $\frac{\log_b n}{2}$, resulting in an expected social welfare of at least $(1 - \frac{1}{n})\frac{2n}{3}\frac{\log_b n}{2} \geq \frac{n \log_b n}{4}$. Moreover, for large enough n, with probability $1 - \frac{1}{n^2}$, the maximum clique is of size at most $4 \log_{\frac{1}{p}} n$ [28]. Hence the expected maximum welfare is at most $n(4 \log_{\frac{1}{1-p}} n + \frac{1}{n^2} n^2) = n(4 \log_{\frac{1}{1-p}} n + 1)$, where the second term bounds the maximum welfare of a partition with the formation of a clique of all agents for the remaining cases occurring with probability at most $\frac{1}{n^2}$. This yields a ratio of about $\frac{1}{16} \log_{\frac{1}{1-p}} b$.

5.2 Random Multipartite Graphs

Consider $k \geq 2$ distinct groups of agents, where the goal is to form diverse coalitions that contain at most one agent from each group. We model this by assigning a negative edge weight of $-n$ to any

[5] We remark that in their model the edges in the cliques occur with probability p whereas they occur with probability $1-p$ in our model.

pair of agents within the same group, rendering them incompatible. Additionally, certain pairs of agents from different groups may also be incompatible. In a random k-partite graph, any pair of agents from different groups is incompatible with probability p. This problem can be formalized as follows.

Let $G = (\{V_1, \cdots, V_k\})$ be a k-partite graph where vertices represent agents. The graph consists of n vertices partitioned into k disjoint "color" classes $V_1, V_2, \ldots, V_k$. All our results hold if k is either a constant or any function satisfying $k = o\left(\frac{n}{\log n}\right)$. Without loss of generality, assume that the color classes are sorted in nonincreasing order by the number of vertices they contain, i.e., $|V_1| \geq |V_2| \geq \cdots \geq |V_k|$.

A k-partite graph is said to be *balanced* if $|V_k| \geq q|V_1|$ holds for some constant $q \in (0, 1)$. A Turán graph $G = (n, k)$ is a special case of a balanced k-partite graph, where each color class contains the same number of vertices, i.e., for all $i \in [k]$, we require $|V_i| = \frac{n}{k}$. We capture these in our second model of random graphs inducing aversion-to-enemies games.

Definition 5.5. A *random balanced k-partite graph* $G = (\{V_1, \ldots, V_k\}, p)$ is a weighted graph where edge weights are sampled independently as follows: each edge between vertices in two different color classes independently takes a weight of $-n$ with probability p, and a weight of 1 with probability $1 - p$; each edge between vertices of the same color class takes a weight of $-n$ with probability 1. The input parameter $p \in (0, 1)$ is called the *perturbation probability*, and it is allowed to depend on n.

A *random Turán graph* $G = (n, k, p)$ is a random balanced k-partite graph where each color class contains the same number of vertices, i.e., for all $i \in [k]$ we have $|V_i| = \frac{n}{k}$.

The goal is to find a partition of maximum welfare for the case when the input is a random balanced k-partite graph (or a random Turán graph). Note that in the cases where $p = 0$ or $p = 1$ the problem becomes trivial. When $p = 0$, all weights between vertices from different color classes are deterministically positive, and the graph G' induced by 1-edges is a complete k-partite graph. In this case, an optimal partition of a Turán graph consists of $\frac{n}{k}$ coalitions, each containing a unique member from each of the color classes $V_1, \ldots, V_k$. For a general balanced k-partite graph, the welfare-maximizing partition contains $|V_k|$ k-cliques, $|V_k| - |V_{k-1}|$ $(k-1)$-cliques, etc. Conversely, when $p = 1$, then all edges in the graph have weight of $-n$, which implies the maximum welfare is obtained by the singleton partition.

We now establish a straightforward upper bound on the maximum welfare. Recall that G' is the graph obtained by removing all negative edges from G. Since G' is k-partite, the maximum clique size in G' is at most k. Thus, Lemma 5.1 implies the following proposition.

Proposition 5.6. *In a random balanced k-partite graph, the maximum welfare is bounded by $\mathcal{SW}(\pi^*) \leq n(k-1)$.*

In our analysis, both k and p can depend on n. We now present polynomial-time algorithms that compute a constant-factor approximation of social welfare when $p = O\left(\frac{1}{k}\right)$, and a $\log_e n$-approximation when p is a constant for random balanced k-partite graphs. We illustrate our results by providing proofs for the special case of random Turán graphs. Due to space limitations, the extension to random balanced k-partite graphs is deferred to the full version of our paper. They are obtained by employing a reduction to Turán graphs, as sketched at the end of the section.

5.2.1 Low Perturbation Regime in random Turán graphs. Algorithm 1 takes as input a random Turán graph, and an accuracy parameter (constant number) $\varepsilon > 0$. In addition, the algorithm takes as input a subset of color classes denoted by $S \subseteq \{V_1, \ldots, V_k\}$. The number of color classes in S is denoted by $k' \leq k$. Our algorithm outputs a partition of the vertices in S into coalitions, meaning that only the vertices in $\bigcup_{V_i \in S} V_i$ are considered, as if the graph consisted of exactly k' color classes, each containing $\frac{n}{k}$ vertices.

Algorithm 1 begins by randomly selecting a vertex to initiate the formation of the first coalition. It then iteratively adds a new vertex w to the coalition if w has only edges of weight 1 towards all current members of the coalition (this ensures there are no $-n$ edges in the created coalitions). This process continues until no additional vertices can be included. Hence, each formed coalition is a maximal clique in the subgraph G' obtained by removing all negative edges, and we refer to these coalitions as maximal cliques. If the size of the resulting maximal clique exceeds $k'\sqrt{1-\varepsilon}$, the vertices in the clique are removed from the pool of available vertices, and the process is repeated with the remaining vertices. However, if at any point the size of the obtained maximal clique is smaller than $k'\sqrt{1-\varepsilon}$, the algorithm terminates and returns the current set of coalitions, with any remaining vertices assigned to singletons coalitions.

Algorithm 1 Greedy coalition formation

Input: $\langle G, S, \varepsilon \rangle$ where $G = (n, k, p)$ is a random Turán graph and $S \subseteq \{V_1, \cdots, V_k\}$ is a subset of color classes with $|S| = k'$.
Output: Partition π on $\bigcup_{V_i \in S} V_i$

$\pi \leftarrow \emptyset$, $R \leftarrow \bigcup_{V_i \in S} V_i$
while $R \neq \emptyset$ **do**
  Select $v \in R$ to begin coalition $C = \{v\}$
  $L \leftarrow R$
  while $\exists w \in L$ with all edges towards C of weight 1 **do**
    $C \leftarrow C \cup \{w\}$, $L \leftarrow L \setminus \{w\}$
  if $|C| \geq k'(\sqrt{1-\varepsilon})$ **then**
    $\pi \leftarrow \pi \cup \{C\}$, $R \leftarrow R \setminus C$
  else
    return $\pi \cup \{\{v\} : v \in R\}$
return π

The following lemma shows that for sufficiently small values of p, by selecting a subset of color classes $S \subseteq \{V_1, \ldots, V_k\}$, where $k' = |S|$, and running Algorithm 1, the algorithm produces nearly $\frac{n}{k}$ maximal cliques, each of which exceeds the size $k'\sqrt{1-\varepsilon}$ with high probability. When $p = O\left(\frac{1}{k}\right)$, the input set S can include all k color classes, i.e., $S = \{V_1, \ldots, V_k\}$. In this case, with high probability, the algorithm finds nearly $\frac{n}{k}$ maximal cliques, each larger than $k\sqrt{1-\varepsilon}$, making the size of these cliques nearly identical to the ideal clique size of k for small values of ε. Consequently, this results in a constant approximation of the social welfare.

Lemma 5.7. *Consider a random Turán graph $G = (n, k, p)$, and a nonempty subset of color classes $S \subseteq \{V_1, \cdots, V_k\}$, and $p = O(\frac{1}{k'})$ for $k' = |S|$. For any fixed $\varepsilon \in (0,1)$ and $\alpha \in (0,1)$, Algorithm 1 returns a partition π with $\alpha\frac{n}{k}$ cliques of size at least $k'\sqrt{1-\varepsilon}$ with probability $1 - e^{-\Theta\left(\frac{nk'}{k}\right)}$.*

Proof. We prove that the size of the first $\frac{n}{k}$ maximal cliques exceeds $k'(\sqrt{1-\varepsilon})$ with high probability. Let C denote a clique that is formed during the ith iteration of the while loop, with a current size of t. A color class is said to be *available* if no vertex from that class has been added to C, yet. The probability that C is maximal is $\left(1-(1-p)^t\right)^{(k'-t)(\frac{n}{k}-i)}$, as this represents the probability that none of the $\frac{n}{k} - i$ remaining vertices in each of the $k' - t$ available color classes can be added to C, due to having at least one edge of weight $-n$ with a vertex in C.

Let X denote the number of maximal cliques in the subgraph induced by the color classes in S, where each clique has size at most $t_0 = k'\sqrt{1-\varepsilon}$. As there are $\binom{k'}{t}(\frac{n}{k})^t$ cliques of size t, and since $\binom{k'}{t}(\frac{n}{k})^t \le \binom{k}{t}(\frac{n}{k})^t \le k^t(\frac{n}{k})^t \le n^t$, the following holds:

$$\begin{aligned}
\mathbb{E}[X] &\le \sum_{t=1}^{t_0} n^t \left(1-(1-p)^t\right)^{(k'-t)(\frac{n}{k}-i)} \\
&\le t_0 n^{t_0} \left(1-(1-p)^{t_0}\right)^{(k'-t_0)(\frac{n}{k}-i)} \\
&= t_0 n^{t_0} (1-(1-p)^{t_0})^{n(\frac{k'-t_0}{k} - \frac{(k'-t_0)i}{n})} \\
&\le t_0 n^{t_0} e^{-(1-p)^{t_0}\left[n(\frac{k'-t_0}{k} - \frac{(k'-t_0)i}{n})\right]} \\
&= t_0 e^{t_0 \log_e n} \cdot e^{-(1-p)^{t_0}\left[n(\frac{k'-t_0}{k} - \frac{(k'-t_0)i}{n})\right]}
\end{aligned}$$

where we used $1 - x \le e^{-x}$. Since $t_0 = k'(\sqrt{1-\varepsilon})$ and $p = O(\frac{1}{k'})$, $(1-p)^{t_0} \ge e^{t_0(-p-p^2)} \to c_0$ for some positive constant c_0. Therefore, the expression can be rewritten as:

$$\begin{aligned}
\mathbb{E}[X] &\le t_0 e^{t_0 \log_e n} \cdot e^{c_0\left[-n(\frac{k'-t_0}{k} - \frac{(k'-t_0)i}{n})\right]} \\
&= k'(\sqrt{1-\varepsilon}) e^{k'\sqrt{1-\varepsilon}\log_e n} \cdot e^{c_0\left[-n\left(\frac{k'}{k}(1-\sqrt{1-\varepsilon}) - \frac{k'(1-\sqrt{1-\varepsilon})i}{n}\right)\right]}.
\end{aligned}$$

For any constant $\alpha \in (0,1)$, while the current iteration satisfies $i \le \alpha\frac{n}{k}$, it holds that

$$\mathbb{E}[X] \le k'(\sqrt{1-\varepsilon}) e^{k'(\sqrt{1-\varepsilon})\log_e n} e^{c_0\left[-n\frac{k'}{k}(1-\alpha)(1-\sqrt{1-\varepsilon})\right]}.$$

Let $a_0 = \sqrt{1-\varepsilon}$ and $b_0 = c_0(1-\alpha)(1-\sqrt{1-\varepsilon})$ be two constant numbers,

$$\mathbb{E}[X] \le k' e^{k'[a_0 \log_e n - b_0 \frac{n}{k}]}.$$

Since $k = o\left(\frac{n}{\log n}\right)$, we have that $\mathbb{E}[X]$ tends to zero as n tends to infinity. By Markov's inequality, the probability of having at least one maximal clique of size at most t_0 is at most $k' e^{-\Theta\left(\frac{nk'}{k}\right)}$.

Thus, the probability of exiting the first while loop during the ith iteration, where $i \le \alpha\frac{n}{k}$, is also at most $k' e^{-\Theta\left(\frac{nk'}{k}\right)}$. By a union bound, the probability that the algorithm exits the while loop before $i = \alpha\frac{n}{k}$ is bounded by

$$\alpha\frac{n}{k} k' e^{-\Theta\left(\frac{nk'}{k}\right)} = e^{-\Theta\left(\frac{nk'}{k}\right)}.$$

Therefore, with probability $1 - e^{-\Theta\left(\frac{nk'}{k}\right)}$, the algorithm returns $\alpha\frac{n}{k}$ cliques of size $k'(\sqrt{1-\varepsilon})$. □

We now prove our main theorem for the low perturbation regime. Note that the theorem extends to the case of random balanced k-partite graphs as we show in the full version.

Theorem 5.8. *Consider aversion-to-enemies games given by random Turán graphs $G = (n, k, p)$, where $k \ge 2$ and $p = O(\frac{1}{k})$. Then there is a polynomial-time algorithm that returns a constant-factor approximation to maximum welfare with probability $1 - e^{-\Theta(n)}$.*

Proof. Fix a small $\varepsilon \in (0,1)$, and consider Algorithm 1 for input $\langle G, S = \{V_1, \ldots, V_k\}, \varepsilon\rangle$. Since $p = O\left(\frac{1}{k}\right)$, Lemma 5.7 implies that the algorithm returns $\alpha\frac{n}{k}$ cliques of size at least $k(\sqrt{1-\varepsilon})$ with probability $1 - e^{-\Theta(n)}$, where α is any constant in the range $(0,1)$. Each clique contains at least $k\sqrt{1-\varepsilon}$ agents, and the utility of every agent in such cliques is at least $k(\sqrt{1-\varepsilon}) - 1$. Therefore, the social welfare of the partition returned by the algorithm is bounded as

$$\mathcal{SW}(\pi) \ge \alpha\frac{n}{k} k\sqrt{1-\varepsilon}(k(\sqrt{1-\varepsilon}) - 1) = \alpha nk(1-\varepsilon) - \alpha n\sqrt{1-\varepsilon}.$$

Moreover, Proposition 5.6 implies that $\mathcal{SW}(\pi^*) \le n(k-1) < nk$. Hence,

$$\frac{\mathcal{SW}(\pi)}{\mathcal{SW}(\pi^*)} \ge \alpha(1-\varepsilon) - \frac{\alpha\sqrt{1-\varepsilon}}{k}. \tag{3}$$

Note that ε is a parameter of our choice, and it can be chosen arbitrarily close to zero. In addition, α is a constant with $\alpha \in (0,1)$, meaning that α can be made arbitrarily close to 1. This implies that the approximation factor as bounded in Equation (3) can be arbitrarily close to $1 - \frac{1}{k} \ge \frac{1}{2}$, where we use that $k \ge 2$. □

As we just showed, the approximation factor can be arbitrarily close to $1 - \frac{1}{k}$. Hence, in case that k tends to infinity as n tends to infinity, we obtain nearly optimal partitions for large n.

5.2.2 High Perturbation Regime for Random Turán Graph. We now present a second algorithm, which uses Algorithm 1 as a subroutine, for the case when the perturbation probability is constant, i.e., $p = c$ for some constant $c \in (0,1)$. Algorithm 2 partitions the set of color classes $\{V_1, \cdots, V_k\}$ into $\lceil ck \rceil$ disjoint sets, so that the sizes of the sets differ by at most one. Denote these disjoint sets by $\{S_1, \ldots, S_{\lceil ck \rceil}\}$.[6]

At least $\lfloor ck \rfloor$ of these sets contain $k' = \left\lfloor \frac{1}{c} \right\rfloor$ color classes. Since k' is a constant, it follows that $p = O\left(\frac{1}{k'}\right)$. Each set of colors forms a subproblem, and by applying Algorithm 1 to each subproblem, we obtain $\alpha\frac{n}{k}$ coalitions of size $k'\sqrt{1-\varepsilon}$ with probability $1 - e^{-\Theta(\frac{n}{k})}$, for some fixed constants $\varepsilon \in (0,1)$ and $\alpha \in (0,1)$, as established by Lemma 5.7. By a union bound, the probability that Algorithm 1 returns fewer than $\alpha\frac{n}{k}$ coalitions of size $k'\sqrt{1-\varepsilon}$ in at least one subproblem is at most $\lceil ck \rceil e^{-\Theta(\frac{n}{k})} \le ne^{-\Theta(\frac{n}{k})}$. Since $k = o\left(\frac{n}{\log n}\right)$,

[6] For example, if $k = 11$ and $c = \frac{1}{3}$, one can take $S_1 = \{V_1, V_2, V_3\}$, $S_2 = \{V_4, V_5, V_6\}$, $S_3 = \{V_7, V_8, V_9\}$, and $S_4 = \{V_{10}, V_{11}\}$.

this probability approaches zero as n increases. Therefore, with probability $1 - ne^{-\Theta(\frac{n}{k})}$, all subproblems return at least $\alpha\frac{n}{k}$ coalitions of size $k'\sqrt{1-\varepsilon}$. This leads to the following lemma.

Lemma 5.9. *Let π be the partition returned by Algorithm 2. Then $\mathcal{SW}(\pi) = \Omega(n)$ with probability $1 - ne^{-\Theta(\frac{n}{k})}$.*

Proof. In at least $\lfloor kc \rfloor$ subproblems, the number of color classes is k', and with probability $1 - ne^{-\Theta(\frac{n}{k})}$, all these subproblems return $\alpha\frac{n}{k}$ coalitions of size $k'\sqrt{1-\varepsilon}$ for some constant $\varepsilon \in (0,1)$ and $\alpha \in (0,1)$. The utility of agents in these coalitions is least $k\sqrt{1-\varepsilon} - 1$. Therefore,

$$\mathcal{SW}(\pi) \geq \lfloor kc \rfloor \alpha \frac{n}{k}(k'\sqrt{1-\varepsilon})(k'\sqrt{1-\varepsilon} - 1).$$

Since $k' = \left\lfloor \frac{1}{c} \right\rfloor$ and c is constant, it follows that $\mathcal{SW}(\pi) \geq nc_0$ for some constant c_0. □

Algorithm 2 Dividing into smaller subproblems

Input: $\langle G, \varepsilon \rangle$ where $G = (n, k, p)$ is a random Turán graph and $p = c$ for some constant $c \in (0,1)$
Output: Partition π

1: $\pi \leftarrow \emptyset$
2: Partition $\{V_1, \ldots, V_k\}$ into $\lceil ck \rceil$ disjoint sets $S_1, \ldots, S_{\lceil ck \rceil}$ that differ in size by at most one
3: **for** each group of colors $S \in \{S_1, \cdots, S_{\lceil kc \rceil}\}$ **do**
4: Let π_S be the partition within vertices in color classes of S, after applying Algorithm 1 on input $\langle G, S, \varepsilon \rangle$
5: $\pi \leftarrow \pi \cup \pi_S$

We now bound the maximum welfare. The proof is similar to our arguments for Erdős-Rényi graphs and deferred to the full version.

Lemma 5.10. *When $k = \Omega(\log n)$ and $p = c$ for some constant $c \in (0,1)$, the maximum welfare satisfies $SW(\pi^*) = O(n \log n)$ with probability $1 - \left(\frac{ce}{2\log_e n}\right)^{\frac{2}{c}\log_e n}$.*

We now prove our main result for the high perturbation regime. Again, the theorem extends to random balanced k-partite graphs.

Theorem 5.11. *Consider aversion-to-enemies games given by random Turán graphs $G = (n, k, p)$, where $p = c$ for some constant $c \in (0,1)$. Then there exists a polynomial-time algorithm that returns a partition π which provides a $\log n$-approximation of the maximum welfare with probability $1 - ne^{-\Theta(\frac{n}{k})} - \left(\frac{ce}{2\log_e n}\right)^{\frac{2}{c}\log_e n+1}$.*

Proof. Lemma 5.9 implies that Algorithm 2 returns a partition π where $\mathcal{SW}(\pi) = \Omega(n)$ with probability $1 - ne^{-\Theta(\frac{n}{k})}$. A simple upper bound on the maximum welfare is provided in Proposition 5.6, which implies $\mathcal{SW}(\pi^*) \leq n(k-1)$. If $k = O(\log n)$, this results in $\mathcal{SW}(\pi^*) = O(n \log n)$. However, when $k = \Omega(\log n)$, it does not provide a useful guarantee. Instead, Lemma 5.10 shows that even in this case, $\mathcal{SW}(\pi^*) = O(n \log n)$ with probability $1 - \left(\frac{ce}{2\log_e n}\right)^{\frac{2}{c}\log_e n}$. By a union bound, we have

$$\mathcal{SW}(\pi) \geq \Omega\left(\frac{1}{\log_e n}\right)\mathcal{SW}(\pi^*).$$

with probability $1 - ne^{-\Theta(\frac{n}{k})} - \left(\frac{ce}{2\log_e n}\right)^{\frac{2}{c}\log_e n+1}$. □

Extension to Balanced Multipartite Graphs. We finish the section by discussing our extension. The main idea is a reduction to the case of Turán graphs. Recall that in a balanced k-partite graph, the input has components of comparable sizes, i.e., $|V_k| \geq q|V_1|$ holds for some constant $q \in (0,1)$. We refine the greedy algorithm by forcing the components' sizes to become equal; the crucial observation is that by considering subsets of agents $V_i' \subseteq V_i$ such that $|V_i'| = |V_k|$, we do not affect the edge distribution, i.e., $-n$ edges are still present with probability p and weight 1 edges with probability $1 - p$. Our analysis for both the low and high perturbation regimes proceed by bounding how much the social welfare changes because of the changes in the sizes.

6 CONCLUSION

We have investigated maximizing social welfare in additively separable hedonic games. This is known to be a very hard problem in a worst-case analysis: approximating welfare better than the n-approximation provided by maximum weight matchings faces computational boundaries. We have strengthened the existing approximation hardness to games with bounded valuations, in particular when restricting them to $\{-1, 0, 1\}$.

By contrast, we have carved out various possibilities to obtain better approximation guarantees. In games with nonnegative total value, a randomized polynomial-time algorithm achieves a $O(\log n)$-approximation. Our proof establishes an interesting connection to the correlation clustering literature. Moreover, going beyond worst-case guarantees, we have defined two stochastic models of aversion-to-enemies games, i.e., the games which cause the inapproximability in the first place. In both models, we perform a high probability analysis. The first stochastic model is based on Erdős-Rényi graphs, where we can efficiently compute partitions that approximate social welfare within a constant factor. The second stochastic model is based on balanced multipartite graphs. We distinguish a low and high perturbation regime, in which we can again guarantee a constant and logarithmic approximation, respectively.

Social welfare is a fundamental objective in ASHGs that deserves further attention in future research. A specific open question is to investigate whether efficient approximation algorithms are possible for symmetric ASHGs with valuations of $\{-1, 1\}$, see our discussion after Theorem 4.1. Moreover, considering welfare approximation in suitable classes of random hedonic games might lead to intriguing discoveries. One candidate are random ASHGs with uniformly random valuations [12]. Finally, we restricted attention to symmetric games, which is *not* without loss of generality for aversion-to-enemies games (cf. Footnote 2). Hence, another direction is to consider asymmetric subclasses of ASHGs.

ACKNOWLEDGMENTS

Martin Bullinger is supported by the AI Programme of The Alan Turing Institute, Vaggos Chatziafratis is supported by an UCSC startup grant and a Hellman's fellowship, and Parnian Shahkar is supported by the National Science Foundation (NSF) under grant CCF-2230414.

REFERENCES

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

[2] Haris Aziz, Felix Brandt, and Hans Georg Seedig. 2013. Computing Desirable Partitions in Additively Separable Hedonic Games. *Artificial Intelligence* 195 (2013), 316–334.

[3] Coralio Ballester. 2004. NP-completeness in hedonic games. *Games and Economic Behavior* 49, 1 (2004), 1–30.

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

[5] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation clustering. *Machine Learning* 56 (2004), 89–113.

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

[7] Felix Brandt and Martin Bullinger. 2022. Finding and Recognizing Popular Coalition Structures. *Journal of Artificial Intelligence Research* 74 (2022), 569–626.

[8] Felix Brandt, Martin Bullinger, and Leo Tappe. 2024. Stability Based on Single-Agent Deviations in Additively Separable Hedonic Games. *Artificial Intelligence* 334 (2024), 104160.

[9] Felix Brandt, Martin Bullinger, and Anaëlle Wilczynski. 2023. Reaching Individually Stable Coalition Structures. *ACM Transactions on Economics and Computation* 11, 1–2 (2023), 4:1–65.

[10] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia (Eds.). 2016. *Handbook of Computational Social Choice.* Cambridge University Press.

[11] Martin Bullinger. 2020. Pareto-Optimality in Cardinal Hedonic Games. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS).* 213–221.

[12] Martin Bullinger and Sonja Kraiczy. 2024. Stability in Random Hedonic Games. In *Proceedings of the 25th ACM Conference on Economics and Computation (ACM-EC).* 212.

[13] Martin Bullinger and René Romen. 2023. Online Coalition Formation under Random Arrival or Coalition Dissolution. In *Proceedings of the 31st Annual European Symposium on Algorithms (ESA).* 27:1–27:18.

[14] Katarína Cechlárová and Jana Hajduková. 2002. Computational complexity of stable partitions with B-preferences. *International Journal of Game Theory* 31, 3 (2002), 353–354.

[15] Katarína Cechlárová and Antonio Romero-Medina. 2001. Stability in Coalition Formation games. *International Journal of Game Theory* 29 (2001), 487–494.

[16] Moses Charikar, Venkatesan Guruswami, and Anthony Wirth. 2005. Clustering with qualitative information. *J. Comput. System Sci.* 71, 3 (2005), 360–383.

[17] Moses Charikar and Anthony Wirth. 2004. Maximizing quadratic programs: Extending Grothendieck's inequality. In *Proceedings of the 45th Symposium on Foundations of Computer Science (FOCS).* 54–60.

[18] Erik D. Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica. 2006. Correlation clustering in general weighted graphs. *Theoretical Computer Science* 361, 2-3 (2006), 172–187.

[19] Nameirakpam Dhanachandra, Khumanthem Manglem, and Yambem Jina Chanu. 2015. Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science* 54 (2015), 764–771.

[20] Dinko Dimitrov, Peter Borm, Ruud Hendrickx, and Shao C. Sung. 2006. Simple Priorities and Core Stability in Hedonic Games. *Social Choice and Welfare* 26, 2 (2006), 421–433.

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

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

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

[24] 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.

[25] Michele Flammini, Bojana Kodric, and Giovanna Varricchio. 2022. Strategyproof mechanisms for friends and enemies games. *Artificial Intelligence* 302 (2022), 103610.

[26] Michele Flammini, Gianpiero Monaco, Luca Moscardelli, Mordechai Shalom, and Shmuel Zaks. 2021. On the Online Coalition Structure Generation Problem. *Journal of Artificial Intelligence Research* 72 (2021), 1215–1250.

[27] Martin Gairing and Rahul Savani. 2019. Computing Stable Outcomes in Symmetric Additively Separable Hedonic Games. *Mathematics of Operations Research* 44, 3 (2019), 1101–1121.

[28] Geoffrey R. Grimmett and Colin J. H. McDiarmid. 1975. On colouring random graphs. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 77. 313–324.

[29] Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. 2015. Correlation clustering with noisy partial information. In *Conference on Learning Theory*. PMLR, 1321–1342.

[30] Claire Mathieu and Warren Schudy. 2010. Correlation clustering with noisy input. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA).* 712–728.

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

[32] Shao C. Sung and Dinko Dimitrov. 2010. Computational Complexity in Additive Hedonic Games. *European Journal of Operational Research* 203, 3 (2010), 635–639.

[33] Chaitanya Swamy. 2004. Correlation Clustering: maximizing agreements via semidefinite programming.. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, Vol. 4. 526–527.

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

[35] David Zuckerman. 2006. Linear degree extractors and the inapproximability of max clique and chromatic number. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing (STOC).* 681–690.