Manipulation of k-Coalitional Games on Social Networks

Naftali Waxman[1] , **Sarit Kraus**[1] and **Noam Hazon**[2]
[1]Department of Computer Science, Bar-Ilan University, Israel
[2]Department of Computer Science, Ariel University, Israel
{sarit, vaxmann}@cs.biu.ac.il, noamh@ariel.ac.il

Abstract

In many coalition formation games the utility of the agents depends on a social network. In such scenarios there might be a manipulative agent that would like to manipulate his connections in the social network in order to increase his utility. We study a model of coalition formation in which a central organizer, who needs to form k coalitions, obtains information about the social network from the agents. The central organizer has her own objective: she might want to maximize the utilitarian social welfare, maximize the egalitarian social welfare, or simply guarantee that every agent will have at least one connection within her coalition. In this paper we study the susceptibility to manipulation of these objectives, given the abilities and information that the manipulator has. Specifically, we show that if the manipulator has very limited information, namely he is only familiar with his immediate neighbours in the network, then a manipulation is almost always impossible. Moreover, if the manipulator is only able to add connections to the social network, then a manipulation is still impossible for some objectives, even if the manipulator has full information on the structure of the network. On the other hand, if the manipulator is able to hide some of his connections, then all objectives are susceptible to manipulation, even if the manipulator has limited information, i.e., when he is familiar with his immediate neighbours and with their neighbours.

1 Introduction

Coalition formation is one of the fundamental research problems in multi-agent systems [Chalkiadakis *et al.*, 2011]. Broadly speaking, coalition formation is concerned with partitioning a population of agents into disjointed teams (or coalitions) with the aim that some system-wide performance measure is maximized. Indeed, in many coalition formation games there is a central organizer that would like to maximize some objective.

One assumption that is usually made is that the utility function of the agents is known and given as an input. However, in some real-world scenarios the organizer obtains the information regarding the utility function directly from the agents. For example, when dividing students into classes, it is a common practice to ask the students about their social relationships [Alon, 2019], since a student is more satisfied if the number of friends she has within the class to which she is assigned is maximized. Similarly, when assigning workers to tasks, a manager would be interested in the interpersonal relationships between potential team members. Ideally, the agents would report their true social relationships so that the organizer will be able to choose the most appropriate coalition structure. However, there might be scenarios in which an agent is better off manipulating the organizer by misreporting his relationships.

Indeed, the problem of manipulation in the context of coalitional games has been studied recently [Wright and Vorobeychik, 2015; Flammini *et al.*, 2017]. These studies have looked for strategyproof mechanisms for forming the coalitions, at the cost of non-optimal social welfare (SW). In this paper we propose a complementary approach. We study in which situations there might be an agent with an incentive to manipulate the organizer, and in which situations no agent has an incentive to manipulate the organizer, and thus a special strategyproof mechanism is not needed (see [Vallée *et al.*, 2014] for a similar approach). This analysis is in the same vein as the works of [Gibbard, 1973] and [Satterthwaite, 1975] in the context of voting, that studied in which situations there might be a voter with an incentive to misreport her true vote.

We focus on k-coalitional games, where exactly k coalitions must be formed [Sless *et al.*, 2018]. We assume that the agents' utilities depend on a social network that represents the social relationships among the agents. Specifically, the social network is modeled as an unweighted graph where the vertices are agents and the edges indicate friendship among the agents. The utility function of an agent is the number of friends she has within the coalition to which she is assigned. Actually, our model is a special case of simple Additively Separable Hedonic Games (ASHGs) [Bogomolnaia *et al.*, 2002]. In addition, there is an organizer that would like to maximize some objective function, thus she needs to obtain the structure of the social network from the agents' reports regarding their friendships. In such situations, it is possible that one manipulative agent would like to misreport his friendship connections, in order to increase his utility. In particular, a

	Directed		*Undirected*	
	Add	*Remove*	*Add*	*Remove*
Max-Util	Strict(F 3a)	Strict(F 2b)	Strict(F 3a)	Strict(F 2a)*
Max-Egal	Strategyproof(T 2)	Strict(F 2e)	LB,UB(F 3b,3c), W-proof(T 1)	Strict(F 2f)
At-least-1	Strategyproof(T 3)	LB(F 2g), UB-Proof(T 4)	Strict(F 3d)	LB(F 2g),UB-proof(T 4)

Table 1: Summary of the results. The parentheses near a result refer to the corresponding figure (F) or theorem (T). The results hold for both full information and distance 2, except for the result with the * , which holds only for the full information case. Key: LB/UB/Strict = the objective is subject to LB/UB/Strict-improvement, W-proof = the objective is weak-proof.

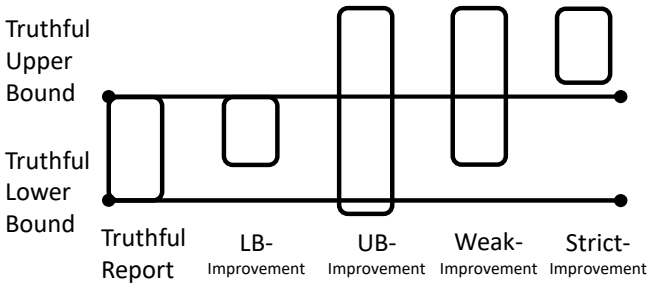


Figure 1: The possible utility values of the manipulator for each manipulation type.

manipulator may hide some of his connections or he may add connections by reporting fake connections (with agents with which the manipulator does not have real connections).

Within these settings we study different objectives for the organizer and analyze their susceptibility or resistance to manipulation. Specifically, we study the objective of maximizing the utilitarian social welfare (Max-Util), maximizing the egalitarian social welfare (Max-Egal), and the At-least-1 objective, where the organizer is only interested in guaranteeing that every agent will have at least one friendship connection within her coalition. Moreover, we study different settings regarding the abilities and information that the manipulator has. Specifically, we study a manipulator that is able to report fake friendship connections (i.e. add edges) and a manipulator that is able to hide some of his friendship connections (i.e., remove edges). We study the situation in which the manipulator has full information regarding the structure of the network and situations in which the manipulator has limited information: he may be familiar only with his connections to his immediate neighbours in the social network (denoted distance 1), or he may also be familiar with their connections to other agents (denoted distance 2). (In both scenarios the manipulator knows about the existence of all agents in the social network, but not how they are connected.)

Table 1 summarizes our results for the full information and distance 2 settings. Overall, all of the objectives are susceptible to manipulation by removing edges, even in the case of distance 2. On the other hand, in some settings there are objectives that are resistant to manipulation by adding edges, even in the case of full information. Note that the results for distance 1 do not appear in the Table, since in almost all of the cases the objectives are resistant to manipulation.

2 Related Work

There are several studies that developed strategyproof mechanisms for forming coalitions. Dimitrov and Sung [2004] discussed ASHGs where agents have both positive and negative edges, and provided a strategyproof algorithm for finding stable outcomes. Rodríguez-Álvarez [2009] analyzed strategyproof core stable solutions' properties. They showed that single lapping rules are necessary and sufficient for the existence of a unique core-stable partition. Aziz *et al.* [2013a] showed that the serial dictatorship mechanism is strategyproof with appropriate restrictions over the agents' preferences. Flammini *et al.* [2017] focused on the utilitarian SW in ASHGs and Fractional Hedonic Games, and proposed strategyproof mechanisms at the cost of non-optimal social welfare. Wright and Vorobeychik [2015] studied a model of ASHG that is very similar to our model, but instead of restricting the number of coalitions they restricted the size of each coalition. Within their model they proposed a strategyproof mechanism that achieves good and fair experimental performance, despite not having a theoretical guarantee. All of these works have looked for strategyproof mechanisms, while our approach is to study in which situations a strategyproof mechanism is indeed needed, and in which situations it is not needed since manipulation is impossible.

Our approach is similar to the approach of Vallée *et al.* [2014], who studied general hedonic games and Sybil attacks, i.e., manipulations, by adding false agents to the game. Vallée *et al.* showed that hedonic games with Nash stability as the solution concept are very robust to Sybil attacks, but when contractual individual stability is the solution concept then every game is manipulable. Recently, Alon [2019] considered the At-least-1 objective, and analyzed whether a group of manipulators can guarantee being in the same coalition in every game. Alon showed that such manipulation is almost always impossible.

3 Definitions

Let $A = \{a_1, \ldots, a_n\}$ be a finite, non-empty set of agents, and let $G = (A, E)$ be a graph with no self loops, representing the friendship connections between the agents. The set of immediate neighbours of a_i in G is denoted by $N(a_i)$. We also refer to G as the *social network*. A coalition $C \subseteq A$ is a subset of agents; we do not require that agents in a coalition form a connected component in the corresponding social network. Let $u(a_i, C)$ be the utility that agent a_i would obtain from being in coalition C. This value is simply the sum

of edges corresponding to the immediate neighbors of a_i that are members of C. That is, $u(a_i, C) = |C \cap N(a_i)|$.

We assume that there is a central organizer that would like to partition the agents into k coalitions in order to satisfy some objective, obj. Let Π_k denote the set of partitions of A that contain exactly k non-empty subsets where $0 < k \leq n$. We refer to elements of Π_k as *coalition structures* (CS), and typically use $P, P', \ldots$ to denote such coalition structures. We assume that the utility of agent a depends only on the members of her coalition. Therefore, if $P \in \Pi_k$, $C \in P$, and $a \in C$, then we use the notation $u(a, P)$ to refer to $u(a, C)$. Note that there may be several coalition structures that satisfy a given objective obj. We denote this set of coalitions as $O_{obj}(G) \subseteq \Pi_k$ and refer to them as *solutions*. In many cases we omit the reference to obj when it is clear.

In our setting the social network is formed based on self reports of the agents. That is, each agent a_i is asked by the organizer to list all of her friendship connections. Formally, let $R = \{r_1, \ldots, r_n\}$ be the set of reports, where $r_i \subseteq A \setminus \{a_i\}$. In such a scenario there might be a manipulative agent m. We begin by assuming that m has full information regarding the social network (we relax this assumption in Section 5) and the objective of the central organizer, obj. Moreover, we assume that m is able to misreport his friendship connections and thus add non-existing edges connecting him to other agents or omit existing edges between him and other agents. We denote these two types of manipulators by m^+ and m^-, respectively. We did not consider a manipulator that is capable of both adding and removing edges since it does not add any new results: in all of our objectives only one capability is needed to show susceptibility to manipulation. Let $G^m = (A, E^m)$ be the resulting social network known to the organizer after the manipulation r_m, and let $N^m(a)$ be the set of immediate neighbours of a in G^m. Note that if G is directed then we assume that m is able to add or remove only outgoing edges, i.e., add an edge $(m, a_i) \notin E$ or remove an edge $(m, a_i) \in E$. m is not able to add or remove incoming edges, i.e., add an edge $(a_i, m) \notin E$ or remove an edge $(a_i, m) \in E$. That is, $N(a) = N^m(a)$ for every agent $a \neq m$. If G is undirected then a manipulation by adding edges is relevant when the organizer adds an edge (a_i, a_j) to G if either $a_i \in r_j$ or $a_j \in r_i$. On the other hand, a manipulation by removing edges is relevant when the organizer adds an edge (a_i, a_j) to G only when both $a_i \in r_j$ and $a_j \in r_i$. When the context is clear we will sometimes refer to G^m as the manipulation.

Clearly, the goal of the manipulator is a successful manipulation. Indeed, in our setting there are several ways to define what a successful manipulation is, since there may be several coalition structures that satisfy obj in G and in G^m, but the utility of m might be different in each such coalition structure. Formally, given a network G and objective obj:

Definition 1. *A manipulation r_m is a* lower bound improvement *(LB-improvement) for a manipulator m if:*

$$\min_{P \in O_{obj}(G^m)} (u(m, P)) > \min_{P \in O_{obj}(G)} (u(m, P)).$$

A manipulation r_m is an upper bound improvement *(UB-improvement) for a manipulator m if:*

$$\max_{P \in O_{obj}(G^m)} (u(m, P)) > \max_{P \in O_{obj}(G)} (u(m, P)).$$

That is, LB-improvement eliminates coalition structures with low utility for the manipulator, while UB-improvement adds coalition structures with higher utility for the manipulator. For example, assume that for an objective obj and a graph G there are two possible CSs. That is, $O_{obj}(G) = \{P_1, P_2\}$. Moreover, assume that $u(m, P_1) = 1$ and $u(m, P_2) = 2$. If there exists a manipulation r_m where $O_{obj}(G^m) = \{P_2\}$ (or any other P satisfying $u(m, P) = 2$) then G^m is a LB-improvement. If there exists a manipulation r_m where $O_{obj}(G^m) = \{P_1, P_3\}$ and $u(m, P_3) = 3$ then G^m is an UB-improvement. LB-improvement can be considered risk aversion of some sort, while UB-improvement suits an optimistic manipulator looking for higher utilities.

There is a stronger variant of manipulation which is both LB- and UB-improvement. An even stronger variant is where every coalition structure is strictly better than every possible coalition structure that would have been generated with m's true preferences. Formally:

Definition 2. *A manipulation r_m is a* weak-improvement *for a manipulator m if it is both LB-improvement and UB-improvement for him. A manipulation r_m is a* strict-improvement *for a manipulator m if:*

$$\min_{P \in O_{obj}(G^m)} (u(m, P)) > \max_{P \in O_{obj}(G)} (u(m, P)).$$

Revisiting our example, a manipulation where $O_{obj}(G^m) = \{P_3\}$ is a strict-improvement. Note that the utility $u(m, P)$ is always calculated over the original graph G with the manipulator's true neighbours. We refer to the different manipulations: LB, UB, weak, and strict-improvement, as *manipulation types*. Finally, we define the susceptibility and resistance to a manipulation type of a given objective.

Definition 3. *An objective obj is* subject to LB-improvement *by manipulator m over (un)directed networks if there exists a (un)directed social network G and a manipulation r_m such that r^m is a LB-improvement for m. Otherwise, we say that obj is* LB-proof *against m.*

The definitions for the other manipulation types are similar. When an objective is both LB- and UB-proof, we say that it is *strategyproof*. Figure 1 demonstrates the possible utility values of the manipulator for each manipulation type.

4 Full Information

We begin our analysis of the objectives and their susceptibility or resistance to the different types of manipulation. To show susceptibility to manipulation, we provide figures that depict the scenarios in which manipulation is possible. We use $k = 2$ in all of our proofs, but they can easily be extended for any k. We use the following notations: In all of the figures the vertex m represents the manipulator. A node is represented by a circle, and a rectangle with a number X represents a clique of X agents. An edge going to (from) a clique represents edges going to (from) all nodes in the clique. An edge going to (from) a clique with a number X represents X edges going to (from) arbitrarily chosen X nodes in the clique. If the graph is directed, then an undirected edge (a, b) represents two directed edges, (a, b) and (b, a). If we prove

a result regarding an undirected graph and refer to a figure with a directed graph then every directed edge represents an undirected edge. Overall, Figure 3 provides scenarios for m^+ and Figure 2 provides scenarios for m^-. Therefore, the dotted edges in Figure 2 are the fake edges that are added by the manipulator, while the dotted edges in Figure 2 are the edges that are removed by the manipulator.

We note that almost all of the susceptibility results in the full information setting (except for Proposition 1) are derived by the results of susceptibility in the distance 2 setting (see Section 5.1). Therefore, in this section we mostly provide the results regarding resistance to manipulation.

4.1 Max-Util

Maximizing the utilitarian social welfare (Max-Util) is a very common objective in hedonic games [Aziz *et al.*, 2015]. It was also studied from the perspective of graph theory, since finding a CS (with k coalitions) that maximizes the utilitarian SW is equivalent to finding a minimum k-cut [Brânzei and Larson, 2009]. Utilitarian SW is defined as the sum of the utilities of all agents. Formally, it is $\sum_{a \in A} u(a, P)$.

Max-Util is always susceptible to manipulation; in all of the situations that we consider, this objective is subject to strict-improvement. Recall that our susceptibility results are derived from the distance 2 setting. However, there is one situation in which the susceptibility to manipulation in the distance 2 setting is not known , and thus we show that even in this situation Max-Util is subject to strict-improvement.

Proposition 1. *Max-Util is subject to strict-improvement by a manipulator m^- over an undirected network.*

Proof. Consider the network G as depicted in Figure 2a. Recall that $k = 2$. Clearly, the minimum 2-cut is obtained by cutting the upper clique ($\{a, b, c, d, e, f\}$) from the rest of the network, yielding a minimum 2-cut of size 3. The manipulator's utility is thus 5. By removing the dotted edges, the minimum 2-cut is obtained by cutting the lower clique ($\{n, o, p, q, r, s\}$), yielding a minimum cut of size 2. The manipulator's utility is strictly improved from 5 to 6. □

4.2 Max-Egal

We now consider the objective of maximizing the egalitarian social welfare (Max-Egal), i.e., maximizing the utility of the agent that is worst off. Formally, it is $\min_{a \in A}(u(a, P))$. The objective egalitarian social welfare has also been studied in ASHGs [Peters, 2016; Aziz *et al.*, 2013b]. Maximizing the egalitarian SW might result in a decrease in the average utility of the agents (which is correlated to the utilitarian SW) but it tries to ensure that all of the agents will have some minimum utility. Now, let $Eg(P, G)$ be the egalitarian SW of a coalition structure P in graph G. The following theorems show that Max-Egal is resistant to manipulation by adding edges. The intuition is that by adding edges the manipulator is not able to pretend to be the agent with the minimum utility, and he may increase the utility of the other agents by at most 1.

Theorem 1. *Max-Egal is weak-proof against manipulator m^+ over undirected networks.*

Proof. Let

$$u_0 = \min_{P \in O(G)}(\{u(m, P)\}), u_1 = \max_{P \in O(G)}(\{u(m, P)\}).$$

We will refer to the CS yielding u_0 as P_0. Assume by contradiction that Max-Egal is subject to weak-improvement. That is, there exists a manipulation r_m and a CS $P^m \in O(G^m)$ such that $u(m, P^m) > u_1$. That is, $P^m \notin O(G)$. In addition,

$$\forall\, P \in O(G^m), u(m, P) > u_0. \tag{1}$$

Since the manipulator can only add edges, it holds that $Eg(P^m, G^m) \geq Eg(P_0, G)$. Moreover, if $Eg(P^m, G^m) = Eg(P_0, G)$ then $P_0 \in O(G^m)$, which is not possible according to inequality 1. Therefore, $Eg(P^m, G^m) > Eg(P_0, G)$. Since the manipulator is able to add at most one new edge to every agent and G is undirected, then $\forall a \in A \setminus \{m\}$,

$$u(a, P^m) \geq Eg(P^m, G^m) - 1 \geq Eg(P_0, G).$$

In addition, $u(m, P^m) > u_1 \geq Eg(P_0, G)$. Overall, $\forall a \in A, u(a, P^m) \geq Eg(P_0, G)$. That is, $Eg(P^m, G) \geq Eg(P_0, G)$, and thus $P^m \in O(G)$, which is a contradiction. □

Theorem 2. *Max-Egal is strategyproof against a manipulator m^+ over directed networks.*

Proof. Let

$$u_0 = \min_{P \in O(G)}(\{u(m, P)\}), u_1 = \max_{P \in O(G)}(\{u(m, P)\}).$$

We will refer to the CS yielding u_0 as P_0. Note that for every $P \in O(G)$ it holds that $Eg(P, G) \leq u_0$. Assume by contradiction that Max-Egal is subject to UB-improvement. That is, there exists a manipulation r_m and a CS $P^m \in O(G^m)$ such that $u(m, P^m) > u_1$. That is, $P^m \notin O(G)$.

Since the manipulator can only add edges, it holds that $Eg(P^m, G^m) \geq Eg(P_0, G)$. Moreover, if $Eg(P^m, G^m) = Eg(P_0, G)$ then $P_0 \in O(G^m)$, which is not possible. Therefore, $Eg(P^m, G^m) > Eg(P_0, G)$. Recall that in directed networks the utility of the other agents does not change. Therefore $\forall a \in A \setminus \{m\}$,

$$u(a, P^m) \geq Eg(P^m, G^m) > Eg(P_0, G).$$

In addition, $u(m, P^m) > u_1 \geq Eg(P_0, G)$. Overall, $\forall a \in A, u(a, P^m) \geq Eg(P_0, G)$. That is, $Eg(P^m, G) \geq Eg(P_0, G)$, and thus $P^m \in O(G)$, which is a contradiction.

Now, assume by contradiction that Max-Egal is subject to LB-improvement. That is, there exists a manipulation r_m such that

$$\forall\, P \in O(G^m), u(m, P) > u_0. \tag{2}$$

That is $P_0 \notin O(G^m)$. Denote an arbitrary CS in $O(G^m)$ as P^m. It holds that $Eg(P^m, G^m) > Eg(P_0, G^m)$ and $Eg(P_0, G) \geq E(P^m, G)$.

Again, in directed networks the utility of the other agents does not change. Therefore, if after the manipulation $Eg(P^m, G^m) > Eg(P^m, G)$, it can only change by the utility of m. But $u(a, P^m) > u(a, P)$, hence even before the manipulation $Eg(P^m, G^m) > Eg(P^m, G)$, in contradiction. □

4.3 At-least-1

In the At-least-1 objective the organizer is only interested in ensuring that every agent will have a utility of at least 1. This objective is very general, and it may result in many possible CSs. It has mostly been studied in the context of graph theory [Stiebitz, 1996; Alon, 2006; Bang-Jensen *et al.*, 2016].

Note that there are some instances where there is no CS that guarantees a utility of at least 1 to every agent. We call such an instance *infeasible*, and we then write $O(G) = \emptyset$. In infeasible instances we assume that the utility of all of the agents is 0. We show that, in contrast to the previous objectives, At-least-1 is less susceptible to manipulations. Specifically, we show that an UB-improvement is almost always impossible, and LB-improvement is impossible by adding directed edges. The intuition is that adding edges is beneficial only if the network is undirected and the new edges transform an infeasible instance into a feasible instance, and by removing edges the manipulator is not able to introduce new solutions.

Theorem 3. *At-least-1 is strategyproof against manipulator m^+ over directed networks.*

Proof. Let $u_1 = \max_{P \in O(G)}(\{u(m, P)\})$, and recall that if $O(G) = \emptyset$ then $u_1 = 0$. Assume by contradiction that the At-least-1 objective is subject to UB-improvement. That is, there exists a manipulation r_m and a coalition structure $P^m \in O(G^m)$ such that $u(m, P^m) > u_1$. That is $P^m \notin O(G)$, and $\forall a \in A, |N^m(a)| \geq 1$. Since $u_1 \geq 0$ then $u(m, P^m) \geq 1$. In addition, recall that in a directed network, $\forall a \in A \setminus \{m\}, N(a) = N^m(a)$, thus $\forall a \in A \setminus \{m\}, u(a, P^m) \geq 1$. Overall, $\forall a \in A, u(a, P^m) \geq 1$ and thus $P^m \in O(G)$, which is a contradiction.

Regarding LB-improvement, since the manipulator is only able to add edges then $O(G) \subseteq O(G^m)$. Therefore, no LB-improvement is possible if $O(G) \neq \emptyset$. If $O(G) = \emptyset$, then $u_1 = 0$. Now, assume by contradiction that the At-least-1 objective is subject to LB-improvement. That is, there exists a manipulation r_m for which $\min_{P \in O(G^m)}(\{u(m, P)\})$ is at least 1. Since $u_1 = 0$ that would imply an UB-improvement as well, which is impossible (as shown above). □

Theorem 4. *At-least-1 is UB-proof against manipulator m^- over directed and undirected networks.*

Proof. Since m is only able to remove edges then for every manipulation r_m it holds that $O(G^m) \subseteq O(G)$. Therefore, $\max_{P \in O(G)}(\{u(m, P)\}) \geq \max_{P \in O(G^m)}(\{u(m, P)\})$ and no UB-improvement is possible. □

5 Limited Information

We now focus on more realistic settings, in which the manipulator is not familiar with the full structure of the network. Instead, we assume that the manipulator is either familiar only with his immediate neighbours in the network, or he may also be familiar with the neighbours of his immediate neighbours. Within this setting we need to revise our definitions of successful manipulations. Specifically, since the manipulator is familiar only with a partial network, we define suitable safe manipulations. Informally, a safe manipulation is a manipulation in which the manipulator is not worse off in all of the possible completions of the partial network, and there exists at least one completion of the partial network in which the manipulator is better off.

Formally, let $A_0 = \{m\}$. Let $G_1 = (A, E_1)$ be a graph, $E_1 \subseteq E$, where $(u, v) \in E_1$ if either u or v belongs to A_0. Similarly, let $A_1 = \{u : (u, v) \in E_1 \vee (v, u) \in E_1\}$. Let $G_2 = (A, E_2)$ be a graph, $E_2 \subseteq E$, where $(u, v) \in E_2$ if either u or v belongs to A_1. Given $d \in \{1, 2\}$, a *possible network* $\overline{G_d}$ of G_d is a network $\overline{G_d} = (A, \overline{E_d})$ where $\overline{E_d} = E_d \cup E'$ such that if $(u, v) \in E'$ then neither u and v belong to A_{d-1}. We assume that the manipulator is familiar with G_d and the objective obj. We denote the settings in which the manipulator is familiar with G_1 (G_2) by *distance 1* (*distance 2*). Indeed, since the manipulator is always familiar with his immediate neighbours he can still add or remove edges as in the full information setting. However, since the manipulator is only familiar with G_d he needs to consider the effect of his manipulation on every possible network $\overline{G_d}$. Given $\overline{G_d}$ and a manipulation r_m, let $\overline{G_d}^m$ be the possible network after the manipulation r_m. We can now revise our definitions of successful manipulations. Given a partial network G_d and an objective obj:

Definition 4. *A manipulation r_m is a* d-safe lower bound improvement *for a manipulator m if for all possible networks $\overline{G_d}$ of G_d it holds that*

$$\min_{P \in O_{obj}(\overline{G_d}^m)}(u(m, P)) \geq \min_{P \in O_{obj}(\overline{G_d})}(u(m, P)).$$

and for at least one possible network $\overline{G_d}$, r_m is a LB-improvement.

An objective is subject to d-safe LB-improvement *by manipulator m over (un)directed networks if there exists a (un)directed partial network G_d and a manipulation r_m such that r_m is a d-safe LB-improvement for m. Otherwise we say that* obj *is* d-safe LB-proof *against m.*

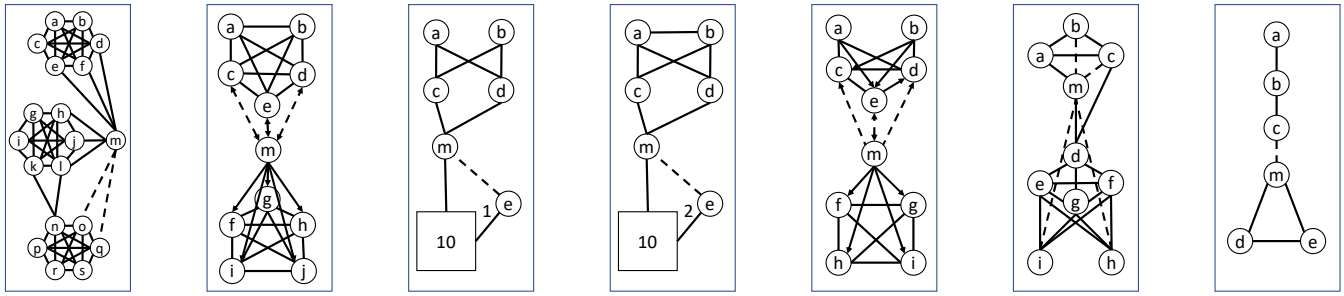
The definitions for the other manipulation types are similar. Note that susceptibility to d-safe manipulation implies susceptibility to d'-safe manipulation for any $d' > d$, as well as to the full information case. Similarly, resistance to manipulation in the full information setting implies resistance to manipulation in the limited information setting.

5.1 Distance 1

We analyze the three objectives and their susceptibility to safe manipulations in the setting of distance 1. Remarkably, even in this setting there are situations where a safe manipulation exists (Proposition 2). However, Theorem 5 shows that for most situations safe manipulation is impossible. The proof is based on extensive enumeration of networks, where we show that either no manipulation exists or there exists only an unsafe manipulation.

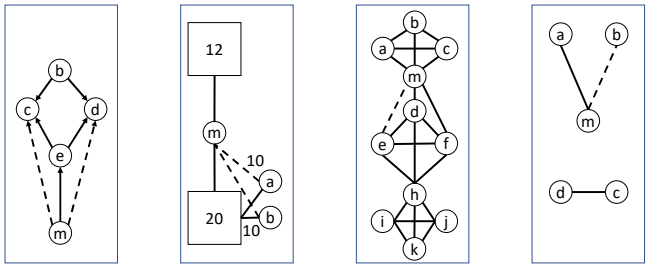
Proposition 2. *Max-Egal against manipulator m^- over directed networks and At-Least-1 against m^+ over undirected networks are subject to 1-safe UB-improvement.*

Proof. Figure 4e provides proof for Max-Egal. Clearly, At-Least-1 is subject to UB-improvement by simply adding any



(a) U/S (b) U/S (c) U/LB (d) U/UB (e) E/S (f) E/S (g) 1/LB

Figure 2: Figures for the proofs of manipulations by m^-. Key:U=Max-Util, E=Max-Egal, 1=At-least-1, S=strict-improvement, UB=UB-improvement, LB=LB-improvement



(a) U/S (b) E/LB (c) E/UB (d) 1/S

Figure 3: Figures for the proofs of manipulations by m^+. Key:U=Max-Util, E=Max-Egal, 1=At-least-1, S=strict-improvement, UB=UB-improvement, LB=LB-improvement

Objective	*Manipulator*	*Network*	*Type*	*Fig*
Max-Util	Add	Both	Strict	3a
Max-Util	Remove	Directed	Strict	2b
Max-Util	Remove	Undirected	LB	2c
Max-Util	Remove	Undirected	UB	2d
Max-Egal	Add	Undirected	LB	3b
Max-Egal	Add	Undirected	UB	3c
Max-Egal	Remove	Directed	Strict	2e
Max-Egal	Remove	Unirected	Strict	2f
At-Least-1	Add	Undirected	Strict	3d
At-Least-1	Remove	Both	LB	2g

Table 2: Summary of susceptibility results for distance 2. Key: LB/UB/Strict = the objective is subject to LB/UB/Strict-improvement.

edge, as it can turn an infeasible instance into a feasible instance. □

Theorem 5. *All three objectives are 1-safe strategyproof except for the cases in Proposition 2.*

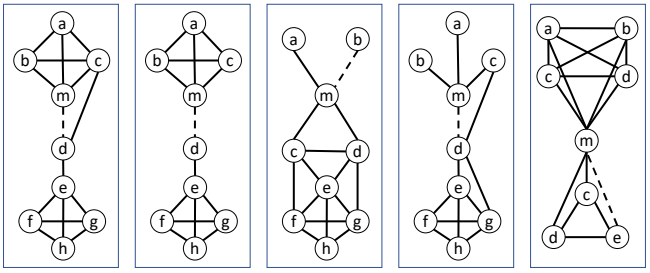
Proof (partial). **Max-Util**. We show that Max-Util is 1-safe strategyproof against manipulators m^- and m^+ over undirected networks. We start with m^-; If $0 < |N(m)| < n-1$, Figures 4a and 4b show that removing any edge is UB-unsafe and LB-unsafe, respectively. Clearly, the graphs in these Figures can be extended to any number of agents and any number of neighbours of the manipulator. For example, to extend figure 4b to arbitrary numbers of agents and neighbours do as following: Connect all of the manipulator's neighbours but one to form a clique (a, b, c in the example). Let the other agents form a clique of their own (e, f, g, h in the example). Lastly connect the manipulator's neighbour which is not in the clique (d in the example) to the other clique with only one edge. This way removing an edge is UB-unsafe.

If $|N(m)| = n-1$, we first show that Max-Util is 1-safe UB-proof. If the manipulator removes only 1 edge then he cannot improve his upper bound at all. If the possible network was a complete graph with n nodes, then removing two edges or more is UB-unsafe. To see that Max-Util is 1-safe LB-proof, look at a complete graph where one edge is missing (an edge not connected to m). Removing any edge in that case can only lower the manipulator's LB.

Continuing with m^+, if $0 < |N(m)| < n-1$ then Figures 4c,4d show that adding any edge is LB- and UB-unsafe respectively. Again, these examples can be extended to fit any number of agents and any number of neighbours the manipulator has. If $|N(m)| = n-1$, the manipulator cannot add edges.

At-Least-1. For At-Least-1, we prove that the objective is 1-safe LB-proof against manipulator m^-. Clearly, At-Least-1 is 1-safe LB-proof over undirected networks, since removing any edge might lead to an infeasible instance. For example, if the manipulator removes an edge from a neighbour that has a degree of one, this neighbour will have a degree of zero in every coalition structure, resulting in an infeasible instance.

Over directed networks, there are two possible cases. If there is an agent a such that $(a, m), (m, a) \in E$, then it is possible that the only feasible coalition structure P is where there is a coalition $C \in P, C = \{m, a\}$. Therefore, removing the edge (m, a) results in an infeasible instance. If there is an agent a such that $(m, a) \in E$ but $(a, m) \notin E$, then it is possible that there is an agent a' such that $(a, a'), (a', m) \in E$ and the only feasible coalition structure p is where there is a coalition $C \in P$, $C = \{m, a, a'\}$. Therefore, removing the edge (m, a) results in an infeasible instance.



(a) R/U/LB (b) R/U/UB (c) A/U/LB (d) A/U/UB (e) R/E/UB

Figure 4: Figures for the proofs of safe and unsafe manipulations, distance 1. Key:U=Max-Util, E=Max-Egal, UB=UB-improvement, LB=LB-improvement, R=Remove, A=Add

Due to space constraints, the complete proof, including the setting of Max-Egal, is provided in the full version of this paper [Waxman *et al.*, 2021]. □

Distance 2

Unlike in the distance 1 setting, the results for the distance 2 setting are almost the same as the results in the full information setting. Indeed, all of the resistance results are derived by the resistance results in the full information setting. Therefore, in this section we provide only susceptibility results summarized in Table 2 where each entry represents a situation, what type of manipulation it is subject to, and a reference to a figure providing a proof. Note that the figures showcase a possible network. The partial network known to the manipulator can easily be derived from them. Overall, we show that, surprisingly, all of the results for the full information setting hold for the distance 2 setting, except for one case: when maximizing the utilitarian SW, with a m^- manipulator over undirected networks. In this case, Proposition 1 shows that the objective is subject to strict-improvement with full information while Figures 2c and 2d only show that it is subject to 2-safe LB- and UB-improvement in the distance 2 setting. Indeed, we believe that the objective is 2-safe weak-proof.

6 Discussion

In this section we discuss our results. We explain the different phenomena that we observe when comparing the results for the different settings. Our results indicate that manipulations over undirected networks are easier than manipulations over directed networks. This is due to the fact that in directed networks the manipulator can only influence his own utility. In undirected networks the manipulator can actually influence the utility of his neighbours, and this additional power enables the manipulation in additional situations. We can also observe that Max-Util is the easiest objective to manipulate when compared with Max-Egal and At-least-1. Indeed, Max-Util is inherently different from the other two objectives: when maximizing the utilitarian SW, the organizer is interested in the average utility, thus the organizer always takes into account the utility of the manipulator. In contrast, in the other two objectives the organizer is interested in the utility of the weakest agents, thus the organizer may not take into account the utility of the manipulator (e.g., if the manipulator already has a utility of 5 in the At-least-1 objective). This characteristic of Max-Egal and At-least-1 objectives can also explain why it is easier to manipulate these objectives by removing edges rather than by adding edges: by removing edges the manipulator can pretend to be the weakest agent, affecting the organizer's choice of the coalition structure. Following this observation, we would expect that manipulating the At-least-1 objective will be harder than manipulating the Max-Egal objective: in At-least-1 the organizer is interested only in ensuring a minimum utility of 1 while in Max-Egal the organizer is also interested in maximizing the minimal utility. However, it turns out that At-least-1 can be manipulated by turning an infeasible instance into a feasible instance, thus there are situations in which At-least-1 is subject to strict-improvement while Max-Egal is weak-proof. Finally, we note that there is a significant difference between results for the settings of distance 1 and distance 2. In the setting of distance 1, the manipulator has very limited knowledge of the network. Therefore, it is hard for the manipulator to estimate the full effect of adding or removing edges, thus finding that a safe manipulation is impossible in most of the situations. Surprisingly, even though the manipulator is not familiar with the full structure of the network in the setting of distance 2, the additional information in this setting is sufficient for finding safe manipulations in many situations.

7 Conclusions and Future Work

We have studied manipulation in the setting of a central organizer that would like to partition a social network into k coalitions. The organizer has a certain objective she would like to satisfy or maximize, whereas there is a manipulator that would like to maximize his utility, i.e., the number of friends within his coalition. We have distinguished between a manipulator who has full information regarding the structure of the social network, and a manipulator that is only familiar with the edges in close proximity to him. An important future research direction to explore is the complexity of finding a manipulation, given a specific objective. We would also like to extend our analysis to the setting with weighted social networks, or social networks with negative edges.

Acknowledgements

This work has been supported in part by the Israel Science Foundation under grant 1958/20, the Ministry of Science, Technology & Space, Israel and the EU project TAILOR under Grant 992215.

References

[Alon, 2006] Noga Alon. Splitting digraphs. *Combinatorics, Probability and Computing*, 15(6):933–937, 2006.

[Alon, 2019] Noga Alon. High school coalitions. https://m.tau.ac.il/~nogaa/PDFS/coalition.pdf, 2019. Accessed: 2019-01-19.

[Aziz *et al.*, 2013a] Haris Aziz, Felix Brandt, and Paul Harrenstein. Pareto optimality in coalition formation. *Games and Economic Behavior*, 82:562–581, 2013.

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

[Aziz *et al.*, 2015] Haris Aziz, Serge Gaspers, Joachim Gudmundsson, Julián Mestre, and Hanjo Taubig. Welfare maximization in fractional hedonic games. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[Bang-Jensen *et al.*, 2016] Jørgen Bang-Jensen, Nathann Cohen, and Frédéric Havet. Finding good 2-partitions of digraphs ii. enumerable properties. *Theoretical Computer Science*, 640:1–19, 2016.

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

[Brânzei and Larson, 2009] Simina Brânzei and Kate Larson. Coalitional affinity games and the stability gap. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[Chalkiadakis *et al.*, 2011] Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. Computational aspects of cooperative game theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(6):1–168, 2011.

[Dimitrov and Sung, 2004] Dinko Dimitrov and Shao Chin Sung. Enemies and friends in hedonic games: individual deviations, stability and manipulation. Discussion paper: 111, Tilburg University, Center for Economic Research, 2004.

[Flammini *et al.*, 2017] Michele Flammini, Gianpiero Monaco, and Qiang Zhang. Strategyproof mechanisms for additively separable hedonic games and fractional hedonic games. In *International Workshop on Approximation and Online Algorithms*, pages 301–316. Springer, 2017.

[Gibbard, 1973] Allan Gibbard. Manipulation of voting schemes: a general result. *Econometrica*, 41(4):587–601, 1973.

[Peters, 2016] Dominik Peters. Graphical hedonic games of bounded treewidth. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[Rodríguez-Álvarez, 2009] Carmelo Rodríguez-Álvarez. Strategy-proof coalition formation. *International Journal of Game Theory*, 38(3):431–452, 2009.

[Satterthwaite, 1975] Mark Allen Satterthwaite. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 10(2):187–217, 1975.

[Sless *et al.*, 2018] Liat Sless, Noam Hazon, Sarit Kraus, and Michael Wooldridge. Forming k coalitions and facilitating relationships in social networks. *Artificial Intelligence*, 259:217–245, 2018.

[Stiebitz, 1996] Michael Stiebitz. Decomposing graphs under degree constraints. *Journal of Graph Theory*, 23(3):321–324, 1996.

[Vallée *et al.*, 2014] Thibaut Vallée, Grégory Bonnet, Bruno Zanuttini, and François Bourdon. A study of sybil manipulations in hedonic games. In *13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014)*, pages 21–28, 2014.

[Waxman *et al.*, 2021] Naftali Waxman, Noam Hazon, and Sarit Kraus. Manipulation of k-coalitional games on social networks. *arXiv preprint arXiv:2105.09852*, 2021.

[Wright and Vorobeychik, 2015] Mason Wright and Yevgeniy Vorobeychik. Mechanism design for team formation. *arXiv preprint arXiv:1501.00715*, 2015.