Multidisciplinary Workshop on Advances in Preference Handling: Papers from the AAAI-14 Workshop

Manipulation and Bribery in Preference Reasoning under Pareto Principle

Ying Zhu and Miroslaw Truszczynski
Department of Computer Science,
University of Kentucky, Lexington, KY 40506, USA

Abstract

Manipulation and bribery have received much attention from the social choice community. We consider these concepts in the setting of preference formalisms, where the *Pareto* principle is used to assign to preference theories *collections* of *optimal* outcomes, rather than a single *winning* outcome as is common in social choice. We adapt the concepts of manipulation and bribery to this setting. We provide characterizations of situations when manipulation and bribery are possible. Assuming a particular logical formalism for expressing preferences, we establish the complexity of determining a possibility for manipulation or bribery. In all cases that do not in principle preclude a possibility of manipulation or bribery, our complexity results show that deciding whether manipulation or bribery are actually possible is computationally hard.

Introduction

In a common *preference reasoning* scenario, a group of agents is presented with a collection of possible *configurations* or *outcomes*. These outcomes come from a *combinatorial* domain, that is, they are characterized by several multivalued attributes and are represented as tuples of attribute values. Each agent has her individual *preferences* on the outcomes. The problem is to *aggregate* these preferences, that is, to define a "group" preference relation or, at the very least, to identify outcomes that could be viewed by the entire group as good consensus choices. This scenario has received much attention in the AI and decision theory communities (Domshlak et al. 2011; Kaci 2011; Lang 2004).

One of the questions it brings up is how to *represent* preferences over a combinatorial domain. A large number of elements in a typical combinatorial domain (exponential in the number of attributes) makes explicit representations impractical. Moreover, with the large number of outcomes to compare and order, it is hardly possible to expect agents to produce orderings accurately capturing their actual preferences. Thus, one resorts to implicit representations which, in order to support both preference elicitation and reasoning, provide concise and intuitive "proxies" to agents' preferences. Often these representations are in terms of sequences of formulas representing a preference order on *properties* of outcomes, with outcomes having most desirable properties being themselves most desirable. For instance, in *answer-set optimization* (Brewka, Niemelä, and Truszczynski 2003), an expression $wine > \neg wine$ is understood as stating the preference for dinners with wine over dinners with any other type of drink (or no drink at all), implicitly defining a preorder on all possible dinners on the menu.[1]

Another key aspect of the scenario above is that of *reasoning*, a fundamental aspect of which is *preference aggregation*. If there is only one user with a single preference, the problem is trivial. But more often than not, an agent has several preferences (for instance not only on the type of drink to take with dinner but also on the appetizer, main dish and dessert selections), or there are several agents, each with her own preference (or preferences). In such cases, to support preference reasoning we *aggregate* preferences into a single consensus preference relation on outcomes or, for some applications, into a set of optimal consensus outcomes.

The problem of preference aggregation is similar to the standard social choice theory scenario (Arrow 1963; Arrow, Sen, and Suzumura 2002). The central objective there is to study methods to aggregate *votes* cast by a group of *voters* into a single *winner* or, in some cases, into a single strict ordering of the candidates. If we think of voters as agents, of candidates as options, and of votes as preferences, the connection between the two areas is evident and, at least to some degree, it has been explored (Chevaleyre et al. 2008). However, the two settings also exhibit some essential differences.

In social choice, the number of options, that is, candidates in an election, is typically small and preferences can be (and are) specified explicitly. Each voter provides her top choice or enumerates all candidates in a strict order of

[1] Several preference representation systems using logic languages to specify preferences have been proposed over the years. The survey by Domshalak et al. (2011), and the monograph by Kaci (2011) discuss several of them. Other popular preference representation formalisms are based on the *ceteris paribus* principle (Boutilier et al. 2004; Wilson 2004) or exploit decision trees (Booth et al. 2010), and often rely on graphical models.

preference. Therefore, the focus is not on languages to represent preferences (votes) but on methods to aggregate them known as *voting rules*. It is required that for each set of votes a voting rule produces a single winner (sometimes a stronger requirement is imposed that a single *strict* ordering of candidates be produced which, in particular, implies a single winner). Most common types of voting rules rely on some form of quantitative scoring (Brams and Fishburn 2002).

In contrast, due to the nature of combinatorial domains, the central problem of preference reasoning is the design of languages to represent preferences. The reasoning task of aggregating preferences is understood as that of defining a semantics for the language — a function that assigns to each *preference theory* (a collection of preferences) a *set* of *preferred* objects from the domain. To identify preferred domain elements, quantitative methods similar to simple voting rules have been considered. However, much of the focus has been on qualitative principles such as *Pareto rule*. The social choice theory research on voting rules has only recently been noted and little effort has been expanded to adapt its research directions and results to the more general setting of preference reasoning.

In this paper we study in the setting of preference reasoning concepts of strategic voting developed in social choice (Gibbard 1973; Satterthwaite 1975; Arrow, Sen, and Suzumura 2002). The two specific problems we consider are *manipulation* and *bribery*. The first problem concerns strategic voting by a voter or a group of voters to secure a better outcome (Gibbard 1973; Satterthwaite 1975). The latter looks into a possibility of securing better outcomes by coercing other voters to vote against their preferences (Faliszewski, Hemaspaandra, and Hemaspaandra 2006). The two problems are clearly relevant to preference reasoning. When a group of agents is to make a decision based on collectively preferred outcomes, understanding whether agents can affect the set of those outcomes in ways that are favorable to them is essential. However, departures from the social choice theory setting make theorems developed there, including the famous Gibbard-Satterthwaite impossibility result concerning manipulation (Gibbard 1973; Satterthwaite 1975) and a slew of results on the complexity of manipulation and bribery under common voting rules (Faliszewski, Hemaspaandra, and Hemaspaandra 2010; Bartholdi, Tovey, and Trick 1989; Faliszewski, Hemaspaandra, and Hemaspaandra 2006; Fitzsimmons, Hemaspaandra, and Hemaspaandra 2013), inapplicable in the setting of preferences over combinatorial domains.

In this work, we model agents' preferences as total *preorders* on the space D of outcomes. That is, we allow indifference among options, not allowed by total orders used as votes in the social choice setting. We select *Pareto efficiency* as the principle of preference aggregation, since it is a common denominator of all preference aggregation techniques considered in preference reasoning. We define the manipulation and bribery problems in this setting, and establish conditions under which manipulation and bribery are possible. In both problems, the key question is whether misrepresenting preferences can improve for a particular agent the quality of the *collection of all preferred outcomes* resulting from preference aggregation.

To be able to decide this question, we have to settle on a way to compare *subsets* of D based on that agent's preference preorder on *elements* of D. This is an interesting and important problem in its own right and has been thoroughly studied (Barberà, Bossert, and Pattanaik 2004). In this paper, we study some commonly used extensions of a total preorder on D to a total preorder on the power set $\mathcal{P}(D)$.

As in the case of manipulation and bribery in social choice, here too, when manipulation or bribery are possible (they often are), the intractability of computing departures from true preferences to improve the outcome for an agent may serve as a barrier against dishonest behaviors. We use our general characterizations to establish the complexity of deciding whether manipulation and bribery are possible when outcomes are subsets of a given set, and a form of preference rules in answer-set optimization (Brewka, Niemelä, and Truszczynski 2003) (that can also be seen as preferences in possibilistic logic (Kaci 2011)) is used to give compact representations of preferences in this domain.

Technical Preliminaries

A *preference* on D is a total preorder on D, that is, a binary relation on D that is reflexive, transitive and total. Each such relation, say $\succeq$, determines two associated relations: *strict preference*, denoted $\succ$, where $x \succ y$ if and only if $x \succeq y$ and $y \not\succeq x$, and *indifference*, denoted $\approx$, where $x \approx y$ if and only if $x \succeq y$ and $y \succeq x$. The indifference relation $\approx$ is an equivalence relation on D and partitions D into equivalence classes, $D_1, \ldots, D_m$, which we always enumerate from the most to the least preferred. Using this notation, we can describe a total preorder $\succeq$ by the expression

$$\succeq: \quad D_1 \succ D_2 \succ \cdots \succ D_m.$$

For example, a total preorder $\succeq$ on $D = \{a, b, c, d, e, f\}$ such that $a \approx d$, $b \approx e \approx f$ and $a \succ b \succ c$ (these identities uniquely determine $\succeq$) is specified by an expression

$$\succeq: \quad a, d \succ b, e, f \succ c.$$

(we omit braces from the notation specifying sets of outcomes to keep the notation simple). For every $a \in D$, we define the *quality degree* of a in $\succeq$, written $q_{\succeq}(a)$, as the unique i such that $a \in D_i$.

Let us consider a group $\mathcal{A}$ of N agents each with her own preference on D. We denote these agents by integers from $\{1, \ldots, N\}$ and their preferences by $\succeq_1, \ldots, \succeq_N$, respectively. We write $D^i_1, \ldots, D^i_{m_i}$ for the equivalence classes of the relation $\approx_i$ enumerated, as above, from the most to the least preferred with respect to $\succeq_i$. We call the sequence $(\succeq_1, \ldots, \succeq_N)$ of preferences of agents in $\mathcal{A}$ a (preference) *profile* of $\mathcal{A}$. For instance,

$$\begin{aligned}
\succeq_1: &\quad a \succ b, c \succ d \succ e, f \\
\succeq_2: &\quad a, c \succ d, e, f \succ b \\
\succeq_3: &\quad f \succ a, c, e \succ b, d.
\end{aligned}$$

is a profile of agents 1, 2 and 3.

Let $\mathcal{A}$ be a set of N agents with a profile $P = (\succeq_1, \ldots, \succeq_N)$. We say that $a \in D$ is *Pareto preferred* in P to $b \in D$ (more formally, Pareto preferred by a group $\mathcal{A}$ of agents with profile P), written $a \succeq_P b$, if for every $i \in \mathcal{A}$, $a \succeq_i b$. Similarly, $a \in D$ is *strictly* Pareto preferred in P to $b \in D$, written $a \succ_P b$, if $a \succeq_P b$ and $b \not\succeq_P a$, that is, precisely when for every $i \in \mathcal{A}$, $a \succeq_i b$, and for at least one $i \in \mathcal{A}$, $a \succ_i b$. Finally, $a \in D$ is *Pareto optimal* in P if there is no $b \in D$ such that $b \succ_P a$. We denote the set of all elements in D that are Pareto optimal in P by $Opt(P)$. Virtually all preference aggregation techniques select "group optimal" elements from those that are Pareto optimal. From now on, we omit the term "Pareto" when speaking about the preference relation $\succeq_P$ on D and optimal elements of D determined by this relation, as we do not consider any other preference aggregation principles.

Let P be the profile given above. Because of the preferences of agents 1 and 3, no outcome can strictly dominate a or f. On the other hand, outcomes b, c, d, e are strictly dominated by a. Thus, $Opt(P) = \{a, f\}$. It is interesting to note that for each of the three agents, the set $Opt(P)$ contains at least one of her "top-rated" outcomes. This is an instance of a general *fairness* property of the Pareto principle.

Theorem 1. *For every profile P of a set $\mathcal{A}$ of agents, and for every agent $i \in \mathcal{A}$, the set $Opt(P)$ of optimal outcomes for P contains at least one most preferred outcome for i.*

Coming back to our example, it is natural to ask how satisfied agent 1 is with the result of preference aggregation and what means might she have to influence the result. If she submits a different ("dishonest") preference, say

$$\succeq: \quad b \succ a \succ c \succ d \succ e, f$$

then, writing P' for the profile $(\succeq, \succeq_2, \succeq_3)$, $Opt(P') = \{a, b, f\}$. It may be that agent 1 would prefer $\{a, b, f\}$ to $\{a, f\}$, for instance, because the new set contains an additional highly preferred outcome for 1. Thus, agent 1 may have an incentive to misrepresent her preference to the group. We will refer to such behavior as *manipulation*. Similarly, agent 1 might keep her preference unchanged but convince agent 3 to replace his preference with

$$\succeq': \quad b \succ f \succ a, c, e \succ d.$$

Denoting the resulting profile $(\succeq_1, \succeq_2, \succeq')$ by P'', $Opt(P'') = \{a, b, f\}$ and, as before, that collection of outcomes may be preferred by agent 1. Thus, agent 1 may have an incentive to try to coerce other agents to change their preference. We will refer to such behavior as *bribery*.

We now formally define *manipulation* and *bribery*. For a profile $P = (\succeq_1, \ldots, \succeq_N)$ and a preference $\succeq$, we write $P_{\succeq_i/\succeq}$ for the profile obtained from P by replacing the preference $\succeq_i$ of the agent i with the preference $\succeq$. Let now $\mathcal{A}$ be a group of N agents with a profile $P = (\succeq_1, \ldots, \succeq_N)$, and let $\succeq'_i$ be a preference of agent i on *subsets* of D.

Manipulation: An agent i can *manipulate* preference aggregation if there is a preference $\succeq$ such that $Opt(P_{\succeq_i/\succeq}) \succ'_i Opt(P)$.

Bribery: An agent t is a target for *bribery* by an agent i, if there is a preference $\succeq$ such that $Opt(P_{\succeq_t/\succeq}) \succ'_i Opt(P)$.

The two concepts closely resemble the corresponding concepts introduced and studied in social choice (Arrow, Sen, and Suzumura 2002; Faliszewski, Hemaspaandra, and Hemaspaandra 2006). The key difference is that in our setting the result of preference aggregation is a *subset* of outcomes and not a single outcome. Thus, when deciding whether to manipulate (or bribe), agents must be able to compare sets of outcomes and not just single outcomes. This is why we assumed that the agent i has a preorder $\succeq'_i$ on $\mathcal{P}(D)$. However, even when D itself is not a combinatorial domain, $\mathcal{P}(D)$ is. Thus, explicit representations of that preorder may be infeasible.

The question then is whether the preorder $\succeq'_i$ of $\mathcal{P}(D)$, which parameterizes the definitions of manipulation and bribery, can be expressed in terms of the preorder $\succeq_i$ on D, as the latter clearly imposes some strong constraints on the former. This problem has received attention from the social choice and AI communities (Barberà, Bossert, and Pattanaik 2004; Brewka, Truszczynski, and Woltran 2010) and it turns out to be far from trivial. The difficulty comes from the fact that there are several ways to "lift" a preorder from D to the power set of D, none of them fully satisfactory (cf. impossibility theorems (Barberà, Bossert, and Pattanaik 2004)). In this paper, we sidestep this issue and simply select and study several most direct and natural "liftings" of preorders on sets to preorders on power sets. We introduce them below. We write X and Y for subsets of D and $\succeq$ for a total preorder on D that we seek to extend to a total preorder on $\mathcal{P}(D)$.

Compare best: $X \succeq^{cb} Y$ if there is $x \in X$ such that for every $y \in Y$, $x \succeq y$.
Compare worst: $X \succeq^{cw} Y$ if there is $y \in Y$ such that for every $x \in X$, $x \succeq y$.

For the next two definitions, we assume that $\succeq$ partitions D into strata $D_1, \ldots, D_m$, as discussed above.

Lexmin: $X \succeq^{lmin} Y$ if for every i, $1 \leq i \leq m$, $|X \cap D_i| = |Y \cap D_i|$, or if for some i, $1 \leq i \leq m$, $|X \cap D_i| > |Y \cap D_i|$ and, for every $j \leq i-1$, $|X \cap D_j| = |Y \cap D_j|$.
Average-rank: $X \succeq^{ar} Y$ if $ar_{\succeq}(X) \leq ar_{\succeq}(Y)$, where for a set $Z \subseteq D$, $ar_{\succeq}(Z)$ denotes the average rank of an element in Z and is defined by $ar_{\succeq}(Z) = \sum_{i=1}^{m} i \frac{|Z \cap D_i|}{|Z|}$.

Manipulation

In this section, we study the manipulation problem in the context of the four extensions of total preorders on D to $\mathcal{P}(D)$. In the case of the *compare best* extension, manipulation is impossible.

Theorem 2. *Let $\mathcal{A}$ be a set of N agents $1, \ldots, N$ with a profile $P = (\succeq_1, \ldots, \succeq_N)$. For every $i \in \mathcal{A}$ and every total preorder $\succeq$, $Opt(P) \succeq_i^{cb} Opt(P_{\succeq_i/\succeq})$.*

This result is a consequence of the fairness property of the Pareto principle stated in Theorem 1. That property implies that set $Opt(P)$ is optimal with respect to the preorder $\succeq_i^{cb}$ on $\mathcal{P}(D)$ among *all* subsets of D. Therefore, it is optimal among all subsets of the form $Opt(P_{\succeq_i/\succeq})$, which implies Theorem 2.[2]

Manipulation is also not possible when the *compare worst* method is used to compare subsets of D.

Theorem 3. *Let $\mathcal{A}$ be a set of N agents $1, \ldots, N$ with a profile $P = (\succeq_1, \ldots, \succeq_N)$. For every $i \in \mathcal{A}$ and every total preorder $\succeq$, $Opt(P) \succeq_i^{cw} Opt(P_{\succeq_i/\succeq})$.*

This time the reason is different but it is again a consequence of the way the Pareto principle works. Let W be the set of all outcomes in $Opt(P)$ that are least preferred for an agent i. To improve the quality of optimal outcomes with respect to her true preference, i would have to submit a dishonest preference that would render all outcomes in W non-optimal. Since preferences of other agents do not change, each such dishonest preference would force into the set of group-optimal outcomes, some that are even worse for i than those in W.

On the other hand, manipulation is possible for every agent using the *lexmin* comparison rule precisely when not every outcome in D is optimal. The reason is that by changing her preference an agent can cause a Pareto-nonoptimal outcome become Pareto-optimal, while keeping the optimality status of every other outcome unchanged.

Theorem 4. *Let $\mathcal{A}$ be a set of N agents $1, \ldots, N$ with a profile $P = (\succeq_1, \ldots, \succeq_N)$ and let $i \in \mathcal{A}$. There exists a total preorder $\succeq$ such that $Opt(P_{\succeq_i/\succeq}) \succ_i^{lmin} Opt(P)$ if and only if $Opt(P) \neq D$.*

For the *average rank* preorder for comparing sets, an agent can manipulate the result to her advantage if there are Pareto-nonoptimal outcomes that are highly preferred by the agent, or when there are Pareto-optimal outcomes that are low in the preference of that agent, as the former can be made optimal and the latter made non-optimal without changing the Pareto-optimality status of other outcomes.

Theorem 5. *Let $\mathcal{A}$ be a set of N agents $1, \ldots, N$ with a profile $P = (\succeq_1, \ldots, \succeq_N)$ and let $i \in \mathcal{A}$. There exists a total preorder $\succeq$ such that $Opt(P_{\succeq_i/\succeq}) \succ_i^{ar} Opt(P)$ if and only if:*

1. *For some $j < ar_{\succeq_i}(Opt(P))$, there exists $a' \in D_j^i$ such that $a' \notin Opt(P)$; or*
2. *For some $j > ar_{\succeq_i}(Opt(P))$, there are $a' \in Opt(P) \cap D_j^i$ and $a'' \in Opt(P)$ such that $a' \neq a''$, $a'' \succeq_k a'$, for every $k \in \mathcal{A}$, $k \neq i$.*

The main message of this section is that when the result of preference aggregation is a set of optimal outcomes, then even the most fundamental and most elementary aggregation rule, Pareto principle, may be susceptible to manipulation. Whether it is or is not depends on how agents measure the quality of a set. If the comparison is based on the best or worst outcomes, manipulation is not possible (a positive result). However, under less simplistic rules such as *lexmin* or *average-rank* the possibility for manipulation emerges (a negative result that we later moderate for some specific preference representation languages by means of the complexity barrier).

[2]Complete proofs of all results can be found at http://www.cs.uky.edu/ai/mani.pdf.

Bribery

In this section, we discuss the bribery problem. Given a set $\mathcal{A}$ of N agents with a profile $P = (\succeq_1, \ldots, \succeq_N)$, the question is whether an agent i can find an agent t, $t \neq i$, and a total preorder $\succeq$ such that $Opt(P_{\succeq_t/\succeq}) \succ_i' Opt(P)$, where $\succeq_i'$ is a "lifted" total preorder that agent i uses to compare subsets of D. Our results on bribery are similar to those we obtained for manipulation, with one notable exception, and show that whether bribery is possible depends on how agents measure the quality of sets of outcomes.

Theorem 6. *Let $\mathcal{A}$ be a set of N agents $1, \ldots, N$ with a profile $P = (\succeq_1, \ldots, \succeq_N)$. For every $i, t \in \mathcal{A}$, $t \neq i$, and every total preorder $\succeq$, $Opt(P) \succeq_i^{cb} Opt(P_{\succeq_t/\succeq})$.*

This result states that when agent i uses best-ranked outcomes in a set as a measure of the quality of that set, then bribery is impossible. No matter which agent t is a target and no matter how that agent changes her preference, the quality of the resulting set of optimal outcomes cannot surpass the quality of the set of outcomes in the original profile.

The situation changes if agents are interested in maximizing the worst outcomes in a set. Unlike in the case of manipulation, the possibility of bribery may now present itself. Given a set $X \subseteq D$ and a total preorder $\succeq$, by $Min_{\succeq}(X)$ we denote the set of all "worst" elements in X, that is the set that contains every element $x \in X$ such that for every $y \in X$, $y \succeq x$.

Theorem 7. *Let $\mathcal{A}$ be a set of N agents $1, \ldots, N$ with a profile $P = (\succeq_1, \ldots, \succeq_N)$ and let $i \in \mathcal{A}$. There exist $t \in \mathcal{A}$, $t \neq i$, and a total preorder $\succeq$ such that $Opt(P_{\succeq_t/\succeq}) \succ_i^{cw} Opt(P)$ if and only if for every $a \in Min_{\succeq_i}(Opt(P))$, there is $a' \in D$ such that $a' \succ_i a$, and $a' \succeq_k a$, for every $k \in \mathcal{A}$, $k \neq t$.*

Bribery is also possible when *lexmin* or *average-rank* methods are used by agents to extend a preorder on D to a preorder on $\mathcal{P}(D)$. Similarly to Theorem 6, the following two theorems are literal generalizations of the earlier results on manipulation.

Theorem 8. *Let $\mathcal{A}$ be a set of N agents $1, \ldots, N$ with a profile $P = (\succeq_1, \ldots, \succeq_N)$ and let $i, t \in \mathcal{A}$, $t \neq i$. There exists a total preorder $\succeq$ such that $Opt(P_{\succeq_t/\succeq}) \succ_i^{lmin} Opt(P)$ if and only if $Opt(P) \neq D$.*

Theorem 9. *Let $\mathcal{A}$ be a set of N agents $1, \ldots, N$ with a profile $P = (\succeq_1, \ldots, \succeq_N)$ and let $i \in \mathcal{A}$. There exist $t \in \mathcal{A}$, $t \neq i$, and a total preorder $\succeq$ such that $Opt(P_{\succeq_t/\succeq}) \succ_i^{ar} Opt(P)$ if and only if:*

1. *For some $j < ar_{\succeq_i}(Opt(P))$, there exists $a' \in D_j^i$ such that $a' \notin Opt(P)$; or*
2. *For some $j > ar_{\succeq_i}(Opt(P))$, there are $a' \in Opt(P) \cap D_j^i$, and $a'' \in Opt(P)$ such that $a' \neq a''$, $a'' \succeq_k a'$, for every $k \in \mathcal{A}$, $k \neq t$.*

Theorems 7, 8 and 9 show that a possibility for bribery may arise when *compare-worst*, *lexmin* and *average-rank* are used to compare sets of outcomes. There is, however, a difference between *lexmin* and the other two methods. For the former, if bribery is possible, then all agents can be targets for bribery (can be used as t in the theorem). This is not the case for the other two methods.

Complexity

So far we studied the problems of manipulation and bribery ignoring the issue of how preferences (total preorders) on D are represented. In this section, we will establish the complexity of deciding whether manipulation or bribery are possible. For this study, we have to fix a preference representation schema.

First, let us assume that preference orders on elements of D are represented explicitly as sequences $D_1, \ldots, D_m$ of the indifference strata, enumerating them from the most preferred to the least preferred. For this representation, the characterizations we presented in the previous section imply that the problems of the existence of manipulation and bribery can be solved in polynomial time. Thus, in the "explicit representation" setting, computational complexity cannot serve as a barrier against them.

However, for combinatorial domains explicit representations are not feasible. We now take for D a common combinatorial domain given by a set U of binary attributes. We view elements of U as propositional variables and assume that each element of U can take a value from the domain $\{true, false\}$. In this way, we can view D as the set of all truth assignments on U. Following a common convention, we identify a truth assignment on U with the subset of U consisting of elements that are true under the assignment. Thus, we can think of D as the power set $\mathcal{P}(U)$ of U.

By taking this perspective, we can use a formula φ over U as a concise implicit representation of the set $M(\varphi) = \{X \subseteq U : X \models \varphi\}$ of all interpretations of U (subsets of U) that satisfy φ, and we can use sequences of formulas to define total preorders on $\mathcal{P}(U)$ $(= D)$.

A *preference statement* over U is an expression

$$\varphi_1 > \varphi_2 > \cdots > \varphi_m, \qquad (1)$$

where all φ_is are formulas over U and $\varphi_1 \vee \cdots \vee \varphi_m$ is a tautology. A preference statement $p = \varphi_1 > \varphi_2 > \cdots > \varphi_m$ determines a sequence $(D_1, \ldots, D_m)$ of subsets of $\mathcal{P}(U)$, where, for every $i = 1, \ldots, m$,

$$D_i = \{X \subseteq U : X \models \varphi_i\} \setminus (D_1 \cup \cdots \cup D_{i-1}).$$

These subsets are disjoint and cover the entire domain $\mathcal{P}(U)$ (the latter by the fact that $\varphi_1 \vee \cdots \vee \varphi_m$ is a tautology). It follows that if $X \subseteq U$, then there is a unique i_X such that $X \in D_{i_X}$. The relation $\succeq_p$ defined so that $X \succeq_p Y$ precisely when $i_X \leq i_Y$ is a total preorder on $\mathcal{P}(U)$. We say that the preference expression p *represents* the preorder $\succeq_p$.[3]

[3]The partition of D into strata that is determined by $\succeq_p$ is not always $(D_1, \ldots, D_m)$ as some sets D_i may be empty.

This form of modeling preferences (total preorders) is quite common. Preference statements were considered by Brewka, Niemelï¿œ and Truszczynski (2003) as elements of preference modules in answer-set optimization programs.[4] Furthermore, modulo slight differences in the notation, preference statements can also be viewed as preference theories of the possibilistic logic (Kaci 2011).

We will now study the complexity of the existence of manipulation and bribery when preferences are given in terms of preference statements. That is, we assume that the input to these problems consists of N preference statements $p_1, \ldots, p_N$. We will denote the total preorders these statements determine by $\succeq_1, \ldots, \succeq_N$, respectively. We will also denote by $(D^i_1, \ldots, D^i_{m_i})$ the sequence of indifference strata determined by p_i, as defined above. We refer to these two problems as the *existence-of-manipulation* (EM) problem and the *existence-of-bribery* (EB) problem, respectively. These problems are parameterized by the method used to compare sets. We denote it by a superscript indicating the method used. Thus, we speak of the EM^{cb} problem (existence of manipulation when *compare-best* method is used), EB^{ar} problem (existence of bribery when *average-rank* method is used), and so on.

Since for the *compare-best* and *compare-worst* methods for comparing sets manipulation is impossible, the EM^{cb} and EM^{cw} problems are (trivially) in P. Similarly, the EB^{cb} is in P, too. For the other cases, we have following results.

Theorem 10. *The EB^{cw} problem is in Δ^P_3 and is both Σ^P_2- and Π^P_2-hard.*

Theorem 11. *The EM^{lmin} and EB^{lmin} problems are NP-complete.*

Theorem 12. *The EM^{ar} and EB^{ar} problems restricted to the case when the agent seeking manipulation or bribery, respectively, has a two-level preference are Σ^P_2-complete.*

The proof of this result follows from Theorems 5 and 9 or, more precisely from simplified characterizations they provide for the case when an agent attempting manipulation or bribery has a two-level preference. Theorem 12 provides a lower bound for the complexity for the the general case. Since both problems are clearly in PSPACE, we obtain the following corollary.

Corollary 1. *The EM^{ar} and EB^{ar} problems are Σ^P_2-hard and in PSPACE.*

Conclusions and Future Work

We studied manipulation and bribery problems in the setting of preference representation and reasoning, where the *Pareto* principle is used for preference aggregation. In this setting, agents submit preferences on elements of the space of outcomes but, when considering manipulation and bribery, they need to assess the quality of *sets* of such elements. In the paper, we considered several natural

[4]The original definition (Brewka, Niemelä, and Truszczynski 2003) allows for more general preference statements. However, they all can be effectively expressed in terms of preference statements as defined here.

ways in which a total preorder on a space of outcomes can be lifted to a total preorder on the space of sets of outcomes. For each of these "liftings", we found conditions characterizing situations when manipulation (bribery) are possible. These characterizations show that for some simple ways to lift preorders from sets to power sets it is impossible for any agent to strategically misrepresent preferences (*compare-best* and *compare-worst* for manipulation, and *compare-best* for bribery). In those cases, the Pareto principle is "strategy-proof".

However, for "more informed" ways to compare sets of outcomes, it is no longer the case. In principle, manipulation and bribery cannot be *a priori* excluded (*lexmin* and *average-rank* for both manipulation and bribery and, interestingly, also *compare-worst* in the case of bribery). To study whether computational complexity may provide a barrier against strategic misrepresentation of preferences, we considered a simple logical preference representation language closely related to possibilistic logic and answer-set optimization. For sets of preferences given in this language and for each way of lifting preorders from sets to power sets for which manipulation and bribery are in some cases possible, we proved that deciding the existence of manipulation or bribery is intractable.

Our work leaves several interesting open problems. First, methods to lift preorders from sets to power sets can be defined axiomatically in terms of properties for the lifted preorders to satisfy. Are there general results characterizing the existence of manipulation (bribery) for lifted preorders specified only by axioms they satisfy? Second, we do not know the exact complexity of the problems EB^{cw}, EM^{ar} and EB^{ar} (the latter two problems are closely related to natural decision problems concerning average weight of models of propositional theories over weighted propositional alphabet and so, may be of more general interest). Finally, there are preference aggregation principles properly extending the Pareto one for which understanding the manipulation and bribery is also of interest.

References

Arrow, K.; Sen, A.; and Suzumura, K., eds. 2002. *Handbook of Social Choice and Welfare*. North-Holland.

Arrow, K. 1963. *Social Choice and Individual Values*. Cowles Foundation Monographs Series. Yale University Press.

Barberà, S.; Bossert, W.; and Pattanaik, P. K. 2004. *Ranking sets of objects*. Springer.

Bartholdi, J.J., I.; Tovey, C.; and Trick, M. 1989. The computational difficulty of manipulating an election. *Social Choice and Welfare* 6(3):227–241.

Booth, R.; Chevaleyre, Y.; Lang, J.; Mengin, J.; and Sombattheera, C. 2010. Learning conditionally lexicographic preference relations. In *Proceedings of ECAI 2010*, 269–274.

Boutilier, C.; Brafman, R. I.; Domshlak, C.; Hoos, H. H.; and Poole, D. 2004. Cp-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH* 21:135–191.

Brams, S., and Fishburn, P. 2002. Voting procedures. In Arrow, K.; Sen, A.; and Suzumura, K., eds., *Handbook of Social Choice and Welfare*. Elsevier.

Brewka, G.; Niemelä, I.; and Truszczynski, M. 2003. Answer set optimization. In *IJCAI*, 867–872.

Brewka, G.; Truszczynski, M.; and Woltran, S. 2010. Representing preferences among sets. In *Proceedings of AAAI 2010*.

Chevaleyre, Y.; Endriss, U.; Lang, J.; and Maudet, N. 2008. Preference handling in combinatorial domains: From ai to social choice. *AI Magazine* 29(4):37–46.

Domshlak, C.; Hüllermeier, E.; Kaci, S.; and Prade, H. 2011. Preferences in AI: An overview. *Artificial Intelligence* 175(7-8):1037 – 1052.

Faliszewski, P.; Hemaspaandra, E.; and Hemaspaandra, L. A. 2006. The complexity of bribery in elections. In *Proceedings of AAAI 2006*, 641–646.

Faliszewski, P.; Hemaspaandra, E.; and Hemaspaandra, L. A. 2010. Using complexity to protect elections. *Commun. ACM* 53(11):74–82.

Fitzsimmons, Z.; Hemaspaandra, E.; and Hemaspaandra, L. A. 2013. Control in the presence of manipulators: Cooperative and competitive cases. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI'13, 113–119. AAAI Press.

Gibbard, A. 1973. Manipulation of voting schemes: A general result. *Econometrica* 41(4):pp. 587–601.

Kaci, S. 2011. *Working with Preferences: Less Is More*. Cognitive Technologies. Springer.

Lang, J. 2004. Logical preference representation and combinatorial vote. *Ann. Math. Artif. Intell.* 42(1-3):37–71.

Satterthwaite, M. A. 1975. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory* 10.

Wilson, N. 2004. Extending cp-nets with stronger conditional preference statements. In *Proceedings of AAAI 2004*, 735–741.