Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Multi-Attribute Proportional Representation

Jérôme Lang
Université Paris-Dauphine, France
lang@lamsade.dauphine.fr

Piotr Skowron
University of Oxford, United Kingdom
p.k.skowron@gmail.com

Abstract

We consider the following problem in which a given number of items has to be chosen from a predefined set. Each item is described by a vector of attributes and for each attribute there is a desired distribution that the selected set should fit. We look for a set that fits as much as possible the desired distributions on all attributes. Examples of applications include choosing members of a representative committee, where candidates are described by attributes such as sex, age and profession, and where we look for a committee that for each attribute offers a certain representation, i.e., a single committee that contains a certain number of young and old people, certain number of men and women, certain number of people with different professions, etc. With a single attribute the problem boils down to the apportionment problem for party-list proportional representation systems (in such case the value of the single attribute is the political affiliation of a candidate). We study some properties of the associated subset selection rules, and address their computation.

1 Introduction

A research department has to choose k members for a recruiting committee. A selected committee should be gender balanced, ideally containing 50% of male and 50% of female. Additionally, a committee should represent different research areas in certain proportions: ideally it should contain 55% of researchers specializing in area A, 25% of experts in area B, and 20% in area C. Another requirement is that the committee should contain 30% junior and 70% senior researchers, and finally, the repartition between local and external members should be kept in proportions 30% to 70 %. The pool of possible members is given in Table 1.

In the given example, if the department wants to select $k = 3$ members, then it is easy to see that there exists no such committee that would ideally satisfy all the criteria. Nevertheless, some committees are better than others: intuitively we feel the sex ratio should be either equal to 2:1 or to 1:2, the area ratio should be equal to 2:1:0, the age ratio to 1:2, and the affiliation ratio to 1:2. Such relaxed criteria can be achieved by selecting Ann, Donna, and George. Now, let us consider the above example for the case when $k = 4$. In such case, the ideal sex ratio should be equal to 2:2, the research area ratio to 2:1:1, the age ratio to 1:3, and the affiliation ratio to 1:3. It can be proved, however, that for $k = 4$ there exists no committee satisfying such relaxed criteria. Intuitively, in such case the best committee is either {Ann, Charlie, Donna, George}, with two externals instead of three, or {Charles, Donna, George, Kevin}, with males being over-represented.

Name	Sex	Group	Age	Affiliation
Ann	F	A	J	L
Bob	M	A	J	E
Charlie	M	A	S	L
Donna	F	B	S	E
Ernest	M	A	S	L
George	M	A	S	E
Helena	F	B	S	E
John	M	B	J	E
Kevin	M	C	J	E
Laura	F	C	J	L

Table 1: An example of the pool of candidates.

In this paper we formalize the intuition given in the above example and define what it means for a committee to be optimal. First, we notice that our model generalizes the *apportionment* problem for proportional representation (Balinski and Young 2001). The central question of the apportionment problem is how to distribute parliament seats between political parties, given the numbers of votes cast for each party: this setting corresponds to our problem when there is a single attribute being the political affiliation of a candidate, and the desired distributions being the proportions of votes cast for different parties.

There is a variety of apportionment methods studied in the literature (see (Balinski and Young 2001) for a survey), evaluated along a set of desirable criteria (Balinski and Young 1979), especially *non-reversal*, *exactness and respect of quota*, *population monotonicity*, and *house monotonicity*. We define the analogs of these properties for multi-attribute domains, and identify which of those are satisfied by our notion of optimal committee.

To emphasize the analogy between our model and apportionment methods, we should provide some discussion on where the desired proportions for attributes come from. Typically, but not always, they come from *votes*. For instance, each voter might give her preferred value for each attribute, and the ideal proportions coincide with the observed frequencies. For instance, out of 20 voters, 10 would have voted for a male and 10 for a female, 13 for a young person and 7 for a senior one, etc. In other contexts, proportions can come from approval ballots. Yet in other contexts, voters can also express their preferred proportions, which are then aggregated. Finally, sometimes, instead of votes, there are "global" preferences on the composition of the committee, expressed directly by the group, imposed by law, or by other constraints that should be respected as much as possible independently of voter preferences.

The multi-attribute case, however, is also substantially different from the single-attribute one. In particular, multi-attribute proportional representation systems exhibit computational problems that do not appear in the single-attribute setting. In the second part of our paper we show that finding an optimal committee is often NP-hard. Yet, we show that this challenge can be addressed by designing efficient approximation and fixed-parameter tractable algorithms.

We believe that the model formalized in this paper has broad applications. As an example, consider a political system where the voters do not vote for the candidates directly, but rather for their opinions on various issues. For instance, quoting (Lang and Xia 2016), in 2012, voters in California had to decide in simultaneous multiple referenda whether to adopt each of the given eleven propositions [1]; a similar vote also took place in Florida. Given that the voters vote on propositions, our algorithms can be used to find a set of candidates that, in some sense, best represents opinions of voters about propositions. The number of propositions can be even larger: for instance, political parties have usually quite elaborate programs in which they refer to tens or hundreds of issues.

As another example, consider a library selecting a set of movies to buy. In ImDB [2] movies can be described by many attributes, such as genre, country, language, year, actors, directors, awards, etc. The users usually look for movies by their attributes. Our algorithms can help such a library finding a representative collection of movies.

After positioning our work with respect to related areas in Section 2, we present our model in Section 3. In Sections 4 and 5 we discuss relevant properties of methods for multi-attribute fair representation. In Section 6 we show that, although computing optimal committees is generally NP-hard, there exist good approximation and fixed-parameter tractable algorithms for finding them. In Section 7 we point to further research issues. *All proofs omitted from the main text are provided in the full version of this paper (Lang and Skowron 2015).*

2 Related work

Our model is related to three distinct research areas:

Voting on multi-attribute domains (see (Lang and Xia 2016) for a survey). There, the aim is to output a single winning combination of attributes (*e.g.*, in multiple referenda, a combination of binary values). Our model in case when $k = 1$ can be viewed as a voting problem on a *constrained* multi-attribute domain (constrained because not all combinations are feasible).

Multiwinner (or committee) elections, and in particular *full proportional representation* (Chamberlin and Courant 1983; Monroe 1995). There, the voters vote directly for candidates and do not consider attributes that characterize them. Thus, in this literature, the term "proportional representation" has a different meaning: these methods are 'representative' because each voter feels represented by some member of the elected committee. The computational aspects of full proportional representation and its extensions have raised a lot of attention lately (Procaccia, Rosenschein, and Zohar 2008; Betzler, Slinko, and Uhlmann 2013; Cornaz, Galand, and Spanjaard 2012; Skowron, Faliszewski, and Slinko 2015; Lu and Boutilier 2013). Our study of the properties of multi-attribute proportional representation is close in spirit to the work of Elkind et al. (Elkind et al. 2014), who gives a normative study of multiwinner election rules. *Budgeted social choice* (Lu and Boutilier 2011) is technically close to committee elections, but it has a different motivation: the aim is to make a collective choice about a set of objects to be consumed by the group (perhaps, subject to some constraints) rather than about the set of candidates to represent voters.

Apportionment for party-list representation systems (see the work of Balinski and Young (Balinski and Young 2001) for a survey). As we already pointed out, the apportionment methods correspond to the restriction of our model to a single attribute (albeit with a different motivation). While voting on multi-attribute domains and multiwinner elections have lead to significant research effort in computational social choice, this is less the case for party-list representation systems. Ding and Lin (Ding and Lin 2014) studied a game-theoretic model for a party-list proportional representation system under specific assumptions, and show that computing the Nash equilibria of the game is hard. Also related is the computation of bi-apportionment (assignment of seats to parties within regions), investigated in a few recent papers (Pukelsheim et al. 2012; Serafini and Simeone 2012; Lari, Ricca, and Scozzari 2014).

Constrained approval voting (CAP) (Brams. 1990; Potthoff 1990) is probably the closest work to our setting (MAPR). In CAP there are also multiple attributes, candidates are represented by tuples of attribute values, there is a target composition of the committee and we try to find a committee close to this target. However, there are also substantial differences between MAPR and CAP. First, in CAP, the target composition of the committee, exogenously defined, consists of a target number of seats *for each combination of attributes* (called a cell), that is, for each $\vec{z} \in D_1 \times \ldots \times D_p$, we have a value $s(\vec{z})$; while in MAPR we have a smaller input consisting of a target number for each value of each attribute. Note that the input in CAP is exponentially large in the number of attributes, which makes it infeasible in practice as soon as this number exceeds a few units (probably CAP was designed for very small numbers

[1] http://en.wikipedia.org/wiki/California_elections,_November_2012

[2] http://www.imdb.com/

of attributes). Second, in CAP, the selection criterion of an optimal committee is made in two consecutive steps: first a set of *admissible committees* is defined, and the choice between these admissible committees is made by using approval ballots, and the chosen committee is the admissible committee maximizing the sum, over all voters, of the number of candidates approved (there is no loss function to minimize as in MAPR). A simple translation of CAP into an integer linear programming problem is given in (Potthoff 1990; Straszak et al. 1993).

3 The model

Let $X = \{X_1, \dots, X_p\}$ be a set of p *attributes*, each with a finite domain $D_i = \{x_i^1, \dots, x_i^{q_i}\}$. We say that X_i is binary if $|D_i| = 2$. We let $D = D_1 \times \dots \times D_p$. Let $C = \{c_1, \dots, c_m\}$ be a set of *candidates*, also referred to as the *candidate database*. Each candidate c_i is represented as a vector of attribute values $(X_1(c_i), \dots, X_p(c_i)) \in D$.[3]

For each $i \leq p$, by π_i we denote a *target distribution* $\pi_i = (\pi_i^1, \dots, \pi_i^{q_i})$ with $\sum_{i=1}^{q_i} \pi_i^j = 1$. We set $\pi = (\pi_1, \dots, \pi_p)$. Typically, n voters have casted a ballot expressing their preferred value on every attribute X_i, and π_i^j is the fraction of voters who have x_i^j as their preferred value for X_i, but the results presented in the paper are independent from where the values π_i^j come from (see the discussion in the Introduction).

The goal is to select a committee[4] of $k \in \{1, \dots, m\}$ candidates (or items) such that the distribution of attribute values is as close as possible to π. Formally, let $S_k(C)$ denote the set of all subsets of C of cardinality k. Given $A \in S_k(C)$, the *representation vector* for A is defined as $r(A) = (r_1(A), \dots, r_p(A))$, where $r_i(A) = (r_i^j(A) | 1 \leq j \leq q_i)$ for each $i = 1, \dots, p$, and $r_i^j(A) = \frac{|\{c \in A : X_i(c) = x_i^j\}|}{k}$. While these representation vectors are normalized, sometimes it will be convenient to use unnormalized vectors, which sum up to k instead of 1. We define $R(A) = (R_1(A), \dots, R_p(A))$, where $R_i(A) = (R_i^j(A) | 1 \leq j \leq q_i)$ for each $i = 1, \dots, p$, and $R_i^j(A) = k.r_i^j(A) = |\{c \in A : X_i(c) = x_i^j\}|$.

Definition 1. *A committee $A \in S_k(C)$ is* perfect *for π if $r_i(A) = \pi_i$ for all i.*

Thus, a perfect committee matches exactly the target distribution. Clearly, there is no perfect committee if for some i, j, π_i^j is not an integer multiplicity of $\frac{1}{k}$. In some of our results we will focus on target distributions such that for each i, j the value $k\pi_i^j$ is an integer. We will refer to such target distributions as to *natural* distributions.

We define metrics measuring how well a committee fits a target distribution, called *loss functions*.

Definition 2. *A* loss function *f maps π and r to $f(\pi, r(A)) \in \mathbb{R}$, and satisfies $f(\pi, r(A)) = 0$ if and only if $\pi = r(A)$.*

There are a number of loss functions that can be considered. As often, the most classical loss functions use L^p norms, with the most classical examples of L^1, L^2, and L^∞. We focus on two representative L^p norms, L^1, and L^∞, but we believe that other choices are also justified and may lead to interesting variants of our model. Consequently, we consider the following loss functions:

- $\| \cdot \|_1 : \|\pi, r(A)\|_1 = \sum_{i,j} |r_i^j(A) - \pi_i^j|$.
- $\| \cdot \|_{1,\max} : \|\pi, r(A)\|_{1,\max} = \sum_i \max_j |r_i^j(A) - \pi_i^j|$.
- $\| \cdot \|_{\max} : \|\pi, r(A)\|_{\max} = \max_{i,j} |\pi_i^j - r_i^j(A)|$.

Now, we are ready to formally define the central problem addressed in the paper.

Definition 3 (OPTIMALREPRESENTATION)**.** *Given X, C, π, k, and a loss function f, find a committee $A \in S_k(C)$ minimizing $f(\pi, r(A))$.*

Example 1. *For the example of the Introduction, we have X = {sex, group, age, affiliation}, $D = \{F, M\} \times \{A, B, C\} \times \{J, S\} \times \{L, E\}$, and $X_1(\text{Ann}) = F$, $X_1(\text{Bob}) = M$ etc. $\{\text{Charlie, Donna, George, Kevin}\}$ is optimal for $\| \cdot \|_1$, with $\|\pi, r(A)\|_1 = 0.5 + 0.1 + 0.1 + 0.1 = 0.8$, and for $\| \cdot \|_{1,\max}$, with $\|\pi, r(A)\|_{1,\max} = 0.4$, but not for $\| \cdot \|_{\max}$. $\{\text{Ann, Charlie, Donna, George}\}$ is optimal for $\| \cdot \|_{\max}$, with $\|\pi, r(A)\|_{\max} = \max(0, 0.2, 0.05, 0.2) = 0.2$, but not for the other criteria.*

4 The single-attribute case

In this section we focus on the single-attribute case ($p = 1$). Without loss of generality, let us assume that the single attribute be party affiliation. Further, let us for a moment assume that for each value x_1^j there are at least k candidates with value x_1^j (this is typically the case in party-list elections). Then finding the optimal committee comes down to apportionment problem for party-list elections, where a fractional distribution s_1 has to be "rounded up" to an integer-valued distribution R_1 such that $\sum_j R_1^j = k$.

There are two main families of apportionment methods: *largest remainders* and *highest average* methods (Balinski and Young 2001). We shall not discuss highest average methods here, because they are weakly relevant to our model. For largest remainders methods, a *quota* q is computed as a function of the number of seats k and the number of voters n. The number of votes for party i is $n_i = n.\pi_i$. The most common choice of a quota is the *Hare quota*, defined as $\frac{n}{k}$; the method based on the Hare quota is called the *Hamilton* method. *One of our aims is to generalize the Hamilton method to multiattribute domains.*

Definition 4 (The largest remainder method.)**.** *The largest remainder method with quota q is defined as follows:*

- *for all i, $s_i^* = \frac{n_i}{q}$ is the ideal number of seats for party i.*
- *each party i receives $s_i = \lfloor s_i^* \rfloor$ seats; let $t_i = s_i - s_i^*$ (called the* remainder*).*
- *the remaining $k - \sum_i s_i$ seats are given to the $k - \sum_i s_i$ parties with the highest remainders t_i.*

[3] By writing $X_j(c_i)$, we slightly abuse notation, that is, we consider X_j both as an attribute name and as a function that maps any candidate to an attribute value; this will not lead to any ambiguity.

[4] We will stick to the terminology "committee" although the meaning of subsets of candidates has sometimes nothing to do with the election of a committee.

Equivalently (see below), the largest remainder methods selects a distribution $(k_1, \ldots, k_q)$ minimizing $\max_{i=1,\ldots,p}(s_i^* - k_i) = \max_{i=1,\ldots,p}(\frac{n_i}{q} - k_i)$, which in the case of Hamilton comes down to minimizing $\max_{i=1,\ldots,p}(\frac{k.n_i}{n} - k_i)$. After defining $\pi_1^i = \frac{n_i}{n}$ for all i, we obtain the following result, that shows that our definition of an optimal committee, with any of the three loss functions, generalize the Hamilton apportionment method.

Proposition 1. *When $p = 1$ and assuming there are at least k items for each attribute, optimal subsets for $\| \cdot \|_1$, $\| \cdot \|_{1,\max}$ and $\| \cdot \|_{\max}$ coincide, and correspond to the subsets given by the Hamilton apportionment method.*

Therefore, our model can be seen as a generalization of the Hamilton apportionment method to more than one attribute. Note that our model can easily extend other largest remainder methods, and our results would be easily adapted. Interestingly, when $p \geq 2$, our three criteria no longer coincide. However, for binary domains, the optimal committee are the same for both loss functions, $\| \cdot \|_1$ and $\| \cdot \|_{1,\max}$, since $\sum_{j=1,2} |r_i^j(A) - \pi_i^j| = 2 \max_{j=1,2} |r_i^j(A) - \pi_i^j|$.

Proposition 2.

1. *For each $p \geq 3$ and binary domains, optimal subsets for $\| \cdot \|_1$ and $\| \cdot \|_{\max}$ may be disjoint, even for $k = 2$.*
2. *For each $p \geq 3$, optimal subsets for $\| \cdot \|_{\max}$ and $\| \cdot \|_{1,\max}$ can be disjoint.*
3. *For each $p \geq 2$, if at least one attribute has 4 values, then optimal subsets for $\| \cdot \|_1$ and $\| \cdot \|_{1,\max}$ can be disjoint.*
4. *For $p = 2$ and binary domains, optimal subsets for $\| \cdot \|_1$ and $\| \cdot \|_{\max}$ may differ.*

These negative results come from the constraints imposed by the candidate database, which prevent the selection on the different attributes to be done independently. In the example of the proof of point 1, for instance, since all items with the value x_2^1 for X_2 have value x_3^1 for X_3, selecting q items with $X_2 = x_2^1$ implies selecting q items with $X_3 = x_3^1$. However, if the database is sufficiently diverse so that no such constraints exist, the optimization can be done separately on each attribute. This is captured by the following notion.

Definition 5. *A candidate database C satisfy the* Full Supply *(FS) property with respect to k if for any $\vec{x} \in D$, C has at least k candidates associated with value vector $\vec{x}$.*

The candidate database of Example 1 does not satisfy FS, even for $k = 1$, because there is not a single candidate with group C and age S. If we ignore attributes *group* and *affiliation*, then we are left with 2 (resp., 3, 2, 3) candidates with value vector FJ (resp. MJ, FS, MS): the reduced database satisfies FS for $k \in \{1, 2\}$.

Proposition 3. *Let (X, C, k) be an optimal committee selection problem. If C satisfies FS w.r.t. k, then the following statements are equivalent:*

- *A is an optimal committee for $\| \cdot \|_1$*
- *A is an optimal committee for $\| \cdot \|_{1,\max}$*
- *for any attribute X_i, A is a Hamilton committee for the single-attribute problem $(\{X_i\}, D^{\downarrow X_i}, \pi_i, k)$, where $D^{\downarrow X_i}$ is the projection of D on $\{X_i\}$.*

Moreover, any $\| \cdot \|_1$ (and $\| \cdot \|_{1,\max}$) optimal committee is optimal for $\| \cdot \|_{\max}$. (The converse does not always hold.)

5 Properties of multi-attribute proportional representation

Several properties of apportionment methods have been studied, starting with (Balinski and Young 1979). We omit their definition in the single-attribute case and directly give their generalizations to our more general model. Let A be any optimal committee for some criterion given π, C and k. We recall that $R_i^j(A) = k\, r_i^j(A)$ denotes the number of elements of A with the attribute X_i equal to x_i^j.

- *Non-reversal*: for any attribute X_i, and attribute values x_i^j, $x_i^{j'}$, if $\pi_i^j > \pi_i^{j'}$ then $r_i^j(A) \geq r_i^{j'}(A)$.
- *Exactness and respect of quota*: for all i, either $R_i^j = \lfloor k\pi_i^j \rfloor$ or $R_i^j = \lceil k\pi_i^j \rceil$.
- *Population monotonicity* (with respect to X_i): consider π and ρ such that there exists j that (a) $\pi_i^j > \rho_i^j$, (b) for all $j', j'' \neq j$, $\frac{\pi_i^{j''}}{\pi_i^{j'}} = \frac{\rho_i^{j''}}{\rho_i^{j'}}$, and (c) for all $i' \neq i$ and all j, $\rho_{i'}^j = \pi_{i'}^j$. Then there is an optimal committee B for ρ such that $r_i^j(A) \geq r_i^j(B)$.
- *House monotonicity*: let B be an optimal committee for π, C and $k' > k$. Then for all i, j, $r_i^j(B) \geq r_i^j(A)$. [5]

In the single-attribute case, it is known for long that the Hamilton method satisfies all these properties except house monotonicity (this failure of house monotonicity is better known under the name *Alabama paradox*).

We start by noticing that if a property fails to be satisfied in the single-attribute case, *a fortiori* it is not satisfied in the multi-attribute case. As a consequence, house monotonicity is not satisfied, even under the FS assumption. We now consider the other properties.

Proposition 4. *Under the full supply assumption, non-reversal, exactness and respect of quota, and population monotonicity are all satisfied, for any of our loss functions. In the general case, non-reversal, exactness and respect of quota are not satisfied. If X_i is a binary variable, and for $\| \cdot \|_1$, population monotonicity with respect to X_i is satisfied; however it is not satisfied in the general case.*

6 Computing Optimal Committees

In this section we now investigate the computation complexity of optimal committees. We start with observing that the problem of deciding whether there is a perfect committee for a given instance is NP-complete.

Proposition 5. *Given set of attributes X, a set of candidates C, a vector of target distributions π, an integer k, deciding whether there is a perfect committee is* NP*-complete.*

[5] Some other properties, such as *consistency*, seem more difficult to generalize to the multi-attribute case. Also, resistance to party merging or party splitting are less relevant in our setting than for political elections and we omit them.

This simple result implies that the decision problem associated with finding an optimal committee (is there a committee whose loss is less than θ?) is NP-hard for *all* loss functions. However, if the number of attributes p is fixed, the problem is solvable in polynomial time.

Proposition 6. *Let p be a constant integer. Given set of p attributes X, a set of candidates C, a vector of target distributions π, an integer k, deciding whether there is a perfect committee is solvable in polynomial time.*

Approximating optimal committees

A natural approach to alleviate the NP-hardness of the problem is to analyze whether it can be well approximated. Before proceeding to presentation of our approximation algorithms, the core technical contribution of this paper, we define the notion of approximability used in our analysis.

Definition 6. *An algorithm $\mathcal{A}$ is an α-additive-approximation algorithm for* OPTIMALREPRESENTATION *if for each instance I of* OPTIMALREPRESENTATION *it holds that $|f(\pi, r(A)) - f(\pi, r(A^*))| \leq \alpha$, where A is the committee returned by $\mathcal{A}$ for I, and A^* an optimal committee.*

It is easy to observe that for binary domains it holds that $\|\pi, r(A)\|_1 = 2\|\pi, r(A)\|_{1,\max}$. This implies that for binary domains, an α-additive-approximation algorithm for $\|\cdot\|_1$ is an $\frac{\alpha}{2}$-additive-approximation algorithm for $\|\cdot\|_{1,\max}$.

In this paper we mostly present computational results for binary domains. However, this assumption is not as restrictive as it may seem—every instance of the OPTIMALREPRESENTATION problem can be transformed to a new instance with binary domains in the following way:

- $X_{\text{new}} = \{X_{ij} \mid i = 1, \ldots, p,\ j = 1, \ldots, |D_i|\}$.
- $C_{\text{new}} = \{c'_l \mid l = 1, \ldots, m\}$.
- $\pi_{\text{new}} = (\pi_{i,j} \mid 1 \leq i \leq p, 1 \leq j \leq |D_i|)$, where for all $i = 1, \ldots, m$, $j = 1, \ldots, p$ and $j = 1, \ldots, |D_i|$, $\pi^0_{i,j} = \pi^j_i$ and $\pi^1_{i,j} = 1 - \pi^j_i$.

The following lemma shows how to obtain approximation guarantees for arbitrary domains having guarantees for the problem transformed to binary domains.

Lemma 1. *For a given committee A and target distribution π, let π_{new} denote target distributions obtained as above and let A_{new} be a committee consisting of such candidates from C_{new} which correspond to the members of A. The following holds:*

1. $\|\pi_{\text{new}}, r(A_{\text{new}})\|_1 = 2\|\pi, r(A)\|_1$.
2. $1 \leq \frac{\|\pi_{\text{new}}, r(A_{\text{new}})\|_{1,\max}}{\|\pi, r(A)\|_{1,\max}} \leq \max_i |D_i|$.
3. $\max(\pi_{\text{new}}, r(A_{\text{new}})) = \max(\pi, r(A))$.

Lemma 1 has interesting implications—first shows that the transformed instance has the same perfect committees as the original instance; then it shows how to obtain additive approximation guarantees for arbitrary domains having guarantees for the problem restricted to binary domains, for different loss functions.

Algorithm 1: Local search approximation algorithm.

```
Parameters:
  $\pi = (\pi_1, \ldots, \pi_p)$—input target distributions.
  $\ell$—the parameter of the algorithm.
$A \leftarrow k$ random items from $C$;
while there exist $C_\ell \subset C$ and $A_\ell \subset A$ such that $|C_\ell| \leq \ell$,
$|A_\ell| \leq \ell$, and $f(\pi, r(A)) > f(\pi, r((A \setminus A_\ell) \cup C_\ell))$ do
    $A \leftarrow (A \setminus A_\ell) \cup C_\ell$;
return $A$;
```

Approximation algorithms

In this section we show an approximation algorithm for the OPTIMALREPRESENTATION problem. The algorithm is given in Algorithm 1 and is parameterized by an integer value ℓ. It starts with a random collection of k samples and, in each step, it looks whether it is possible to replace some ℓ items from the current solution with some other ℓ items to obtain a better solution. The algorithm continues until it cannot find any pair of sets of ℓ items that improves the current solution. As we show now, the approximation guarantees depend on the value of the parameter ℓ.

Theorem 1. *For binary domains natural distributions, and for the $\|\cdot\|_1$ loss function, the local search algorithm defined as Algorithm 1 with $\ell = 1$ is a $|X|$-additive-approximation algorithm for* OPTIMALREPRESENTATION.

Proof. Let A^* denote an optimal solution for a given instance I of the problem of finding a perfect committee. Let $A \in S_k(C)$ denote the set returned by the local search algorithm from Algorithm 1. From the condition in the "while" loop, we know that there exist no $c \in C$ and $a \in A$ such that $\|\pi, r(A)\|_1 > \|\pi, r((A \setminus \{a\}) \cup \{c\})\|_1$. Now, let $X_{\text{ex}} \subseteq X$ denote the set of all attributes for which A achieves exact match with π, that is, such that for each $X_i \in X_{\text{ex}}$, we have that $r^1_i(A) = \pi^1_i$ and $r^2_i(A) = \pi^2_i$.

Let us consider the procedure consisting in taking the items from $A \setminus A^*$ and, one by one, replace them with arbitrary items from $A^* \setminus A$. This procedure, in $|A \setminus A^*|$ steps, transforms A into an optimal solution A^*. We now estimate the total gain g induced by this procedure. For each item $a \in A \setminus A^*$, by $a' \in A^* \setminus A$ we denote the item which was taken to replace a in the procedure. For each attribute $X_i \in X$ we define the gain $g_i(a, a')$ of replacing a by a' as:

$$g_i(a, a') = \sum_{j \in \{1,2\}} \left(|r^j_i(A) - \pi^j_i| - |r^j_i(A \setminus \{a\} \cup \{a'\}) - \pi^j_i| \right).$$

We now extend this definition to sets of k candidates:

$$g_i(B, B') = \sum_{j \in \{1,2\}} \left(|r^j_i(A) - \pi^j_i| - |r^j_i((A \setminus B) \cup B') - \pi^j_i| \right).$$

If $X_i \in X_{\text{ex}}$, then $r_i(A) = \pi_i$, and so the replacement cannot improve the quality of the solution relatively to X_i, hence

$$\sum_{i \in X_{\text{ex}}} g_i(A \setminus A^*, A^* \setminus A) \leq 0.$$

Note that $g_i(a, a') \in \{-\frac{2}{k}, 0, \frac{2}{k}\}$. Moreover, for each attribute $X_i \notin X_{\text{ex}}$ there are two possible cases:

1. $r_i^j(A) > \pi_i^j$ and each exchange of candidate that results in a negative gain increases $r_i^j(A)$.
2. $r_i^j(A) < \pi_i^j$ and each exchange that results in a negative gain decreases $r_i^j(A)$.

Intuitively, 1. and 2. mean that for attributes outside of X_{ex}, the negative gains cumulate. Formally, for each $X \notin X_{\mathrm{ex}}$:

$$g_i(A \setminus A^*, A^* \setminus A) \leq \sum_{a \in A \setminus A^*} g_i(a, a'). \qquad (1)$$

From the condition in the "while" loop, we have that for each $a \in A \setminus A^*$: $\sum_i g_i(a, a') \leq 0$, and so: $\sum_i \sum_{a \in A \setminus A^*} g_i(a, a') \leq 0$.

We now give the following sequence of inequalities:

$$\begin{aligned} g &= \sum_i g_i(A \setminus A^*, A^* \setminus A) \\ &= \sum_{i \in X_{\mathrm{ex}}} g_i(A \setminus A^*, A^* \setminus A) + \sum_{i \notin X_{\mathrm{ex}}} g_i(A \setminus A^*, A^* \setminus A) \\ &\leq \sum_{i \notin X_{\mathrm{ex}}} g_i(A \setminus A^*, A^* \setminus A) \leq \sum_{i \notin X_{\mathrm{ex}}} \sum_{a \in A \setminus A^*} g_i(a, a') \\ &\leq - \sum_{i \in X_{\mathrm{ex}}} \sum_{a \in A \setminus A^*} g_i(a, a') \leq |X_{\mathrm{ex}}| \cdot k \cdot \frac{2}{k} = 2|X_{\mathrm{ex}}|. \end{aligned} \qquad (2)$$

Finally, for each attribute $X_i \in X_{\mathrm{ex}}$ the loss relative to X_i, *i.e.*, $|r_i^0 - \pi^0| + |r_i^1 - \pi^1|$, is at most 2. Thus, we get $g \leq 2(|X| - |X_{\mathrm{ex}}|)$, which leads to $g \leq |X|$. □

Is the bound $|X|$ from Theorem 1 a good result? One way to interpret this result is to observe that a solution that for half of the attributes gives exact match, and for other half is arbitrarily bad, is an $|X|$-approximate solution. We do not know whether the bound $|X|$ is reached, but we now show that a lower bound on the error made by the algorithm with $\ell = 1$ is $\frac{2}{3}|X|$ (see Example 2 in the full version of this paper (Lang and Skowron 2015)).

A better approximation bound can be obtained with $\ell = 2$, however it requires much more involved analysis:

Theorem 2. *For binary domains ($|D_i| = 2$, for each $1 \leq i \leq p$), natural distributions, and for $\|\cdot\|_1$ loss function, the local search algorithm from Algorithm 1 with $\ell = 2$ is a $\frac{\ln(k/2)}{2\ln(k/2)-1}\left(|X| + \frac{6|X|}{k}\right)$-additive-approximation algorithm for* OPTIMALREPRESENTATION.

Since a brute-force algorithm can be used to compute an optimal solution for small values of k, Theorem 2 implies that for every $\epsilon > 0$ we can achieve an additive approximation of $\frac{1}{2}(|X| + \epsilon)$, that is we can guarantee that the solution returned by our algorithm will be at least 4 times better than a solution that is arbitrarily bad on each attribute. A natural open question is whether the local search algorithm achieves even better approximation guarantees for larger values of ℓ.

One may argue that the restriction to normal target distributions is a strong one. However, for a given vector of target distributions π, we can easily find a vector π_N of target normal distributions such that $\|\pi, \pi_N\|_1 \leq \frac{2X}{k}$. Thus, the results from Theorems 1 and 2 can be modified by providing approximation ratio worse by an additive value of $\frac{2X}{k}$ but valid for arbitrary target distributions. Again, since an optimal solution can easily be computed for small values of k, we can get arbitrarily close to the approximation guarantees given by Theorems 1 and 2, even for non-normal target distributions.

Example 3 in the full version of this paper (Lang and Skowron 2015) provides a lower bound of $\frac{2X}{7}$ for the approximation ratio of the local search algorithm from Algorithm 1 with $\ell = 2$.

Parameterized Complexity

In this section, we study the parameterized complexity of the problem of finding a perfect committee. We are specifically interested whether for some natural parameters there exist fixed parameter tractable (FPT) algorithms. We recall that the problem is FPT for a parameter P if its each instance I can be solved in time $O(g(P) \cdot \mathrm{poly}(|I|))$, where g is some arbitrary computable function.

From the point of view of parameterized complexity, FPT is seen as the class of easy problems. There is also a whole hierarchy of hardness classes, $\mathrm{FPT} \subseteq W[1] \subseteq W[2] \subseteq \cdots$. For details, we point the reader to appropriate overviews (Downey and Fellows 1999; Niedermeier 2006; Flum and Grohe 2006).

Obviously, the problem admits an FPT algorithm for the parameter m. Now, we present a negative result for parameter k (committee size) and a positive result for the parameter p (number of attributes).

Theorem 3. *The problem of deciding whether there exists a perfect committee is* W[1]*-hard for the parameter k, even for binary domains.*

Theorem 4. *For binary domains, there is an* FPT *algorithm for the perfect committee problem for parameter p.*

Proof. Each item can be viewed as a vector of values indexed with the attributes; there are 2^p such possible vectors: $v_1, \ldots, v_{2^p}$. For each v_i, let a_i denote the number of items that correspond to v_i. Consider the following integer linear program, in which each variable b_i is the number of candidates corresponding to v_i in a perfect committee.

$$\text{minimize} \sum_{i=1}^{2^p} b_i$$

subject to:

$$\begin{aligned} &\text{(a)} : b_i \geq 0 && 1 \leq i \leq 2^p \\ &\text{(b)} : b_i \leq a_i && 1 \leq i \leq 2^p \\ &\text{(c)} : \sum_{i=1}^{2^p} b_i = k && \\ &\text{(d)} : \sum_{i : v_i[j]=1} b_i = \pi_i^1 && 1 \leq j \leq p \end{aligned}$$

This linear program has 2^p variables, thus, by the result of Lenstra (Lenstra 1983, Section 5) it can be solved in

FPT time for parameter p. Below we provide an example illustrating this proof. □

We conclude this section by a short discussion. Finding an optimal committee is likely to be difficult if the candidate database C is large, and the number of attributes not small. Assume $|C|$ is large compared to the size of the domain $\prod_{i=1}^{p} |D_i|$, that each attribute value appears often enough in C and that there is no strong correlation between attributes in C: then, the larger $|C|$, the more likely C satisfies Full Supply, in which case finding an optimal committee is easy. The really difficult cases are when $|C|$ is not significantly larger than the domain, or when C shows a high correlation between attributes.

7 Conclusion

We have defined, and studied, multi-attribute generalizations of a well-known apportionment method (Hamilton), albeit with motivations that go far beyond party-list elections (such as the selection of a common set of items). We have shown positive and negative results concerning the properties satisfied by these generalizations and their computation, but a lot remains to be done. Note that other largest remainder apportionment methods can be generalized in a similar way, but it is unclear how largest-average methods can be generalized.

Acknowledgements

Many thanks to Eunjung Kim for giving us the initial idea of the paper, and for fruitful discussions. Jérôme Lang is supported by the ANR project CoCoRICo-CoDec.

References

Balinski, M., and Young, P. 1979. Criteria for proportional representation. *Operations Research* 27(1):80–95.

Balinski, M., and Young, P. 2001. *Fair Representation : Meeting the Ideal of One Man One Vote*. Brookings Institution Press, second edition.

Betzler, N.; Slinko, A.; and Uhlmann, J. 2013. On the computation of fully proportional representation. *Journal of Artificial Intelligence Research* 47:475–519.

Brams., S. J. 1990. Computer-assisted constrained approval voting. *Interfaces* 20(5):67–80.

Chamberlin, B., and Courant, P. 1983. Representative deliberations and representative decisions: Proportional representation and the Borda rule. *American Political Science Review* 77(3):718–733.

Cornaz, D.; Galand, L.; and Spanjaard, O. 2012. Bounded single-peaked width and proportional representation. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI-2012)*, 270–275.

Ding, N., and Lin, F. 2014. On computing optimal strategies in open list proportional representation: The two parties case. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI-14)*, 1419–1425.

Downey, R., and Fellows, M. 1999. *Parameterized Complexity*. Springer-Verlag.

Elkind, E.; Faliszewski, P.; Skowron, P.; and Slinko, A. 2014. Properties of multiwinner voting rules. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2014)*, 53–60.

Flum, J., and Grohe, M. 2006. *Parameterized Complexity Theory*. Springer-Verlag.

Lang, J., and Skowron, P. 2015. Multi-attribute proportional representation. Technical Report arXiv:1509.03389 [cs.AI].

Lang, J., and Xia, L. 2016. Voting over multiattribute domains. In Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A., eds., *Handbook of Computational Social Choice*. Cambridge University Press. chapter 9.

Lari, I.; Ricca, F.; and Scozzari, A. 2014. Bidimensional allocation of seats via zero-one matrices with given line sums. *Annals OR* 215(1):165–181.

Lenstra, H. W. 1983. Integer programming with a fixed number of variables. *Mathematics of Operations Research* 8(4):538–548.

Lu, T., and Boutilier, C. 2011. Budgeted social choice: From consensus to personalized decision making. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-2011)*, 280–286.

Lu, T., and Boutilier, C. 2013. Multiwinner social choice with incomplete preferences. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013)*.

Monroe, B. L. 1995. Fully proportional representation. *American Political Science Review* 89:925–940.

Niedermeier, R. 2006. *Invitation to Fixed-Parameter Algorithms*. Oxford University Press.

Potthoff, R. 1990. Use of linear programming for constrained approval voting. *Interfaces* 20(5):79–80.

Procaccia, A.; Rosenschein, J.; and Zohar, A. 2008. On the complexity of achieving proportional representation. *Social Choice and Welfare* 30(3):353–362.

Pukelsheim, F.; Ricca, F.; Simeone, B.; Scozzari, A.; and Serafini, P. 2012. Network flow methods for electoral systems. *Networks* 59(1):73–88.

Serafini, P., and Simeone, B. 2012. Parametric maximum flow methods for minimax approximation of target quotas in biproportional apportionment. *Networks* 59(2):191–208.

Skowron, P.; Faliszewski, P.; and Slinko, A. 2015. Achieving fully proportional representation: Approximability result. *Artificial Intelligence* 222:67–103.

Straszak, A.; Libura, M.; Sikorski, J.; and Wagner, D. 1993. Computer-assisted constrained approval voting. *Group Decision and Negotiation* 2(4):375–385.