Maximizing Index Diversity in Committee Elections

Paula Böhm, Robert Bredereck, Till Fluschnik[1]

Abstract

We introduce two models of multiwinner elections with approval preferences and labelled candidates that take the committee's diversity into account. One model aims to find a committee with maximal diversity given a scoring function (e.g. of a scoring-based voting rule) and a lower bound for the score to be respected. The second model seeks to maximize the diversity given a minimal satisfaction for each agent to be respected. To measure the diversity of a committee, we use multiple diversity indices used in ecology and introduce one new index. We define (desirable) properties of diversity indices, test the indices considered against these properties, and characterize the new index. We analyze the computational complexity of computing a committee for both models with most of the indices considered and scoring functions of well-known voting rules, and investigate the influence of weakening the score or satisfaction constraints on the diversity empirically.

1 Introduction

In the realm of decision-making, where alternatives are chosen from a larger pool of options, considerations of both quality and diversity become crucial. This challenge presents itself in varied contexts, such as project funding and academic committee selections. For instance, consider the scenario faced by a city government engaging in participatory budgeting (PB). Here, residents propose community projects characterized by target groups (e.g., children, seniors) and objectives (e.g., environmental protection, education).[2] Similarly, consider the formation of a university hiring committee, where the aim is to select individuals who are not only seen as experts by the colleagues electing them, but also contribute to a diverse range of perspectives (e.g., scientists, non-scientific staff) or disciplines (e.g., math, physics). In both cases, the task becomes to select alternatives with both a high level of satisfaction for the voters (which we measure by a *score*) and high level of diversity (which we measure by an *index*).

Traditional models, as seen in many previous studies, address this by setting hard quotas or constraints, ensuring a specific representation of specific labels. While such an approach ensures a certain minimum diversity, it often disregards the fluid and multifaceted nature of diversity itself. In particular, when many labels are to be respected, it seems impossible to come up with any meaningful constraints without restricting the possible satisfaction of the voters unpredictably. In contrast, our study introduces two models that incorporate diversity indices rather than strict constraints, while ensuring that either the total score of the committee or the satisfactions of the agents cannot be worsened arbitrarily. This approach, applied to the multiwinner election context, allows for a more nuanced and dynamic assessment of diversity. By employing established ecological indices alongside a newly proposed one, we introduce this elsewhere-established approach into the area of computational social choice.

Our Contributions. We introduce two models for incorporating diversity indices into multiwinner elections with approval preferences and labelled candidates. Our models aim to find a committee with maximal diversity given (1) a scoring function and a lower bound for the score, or (2) a minimal satisfaction for each agent. To measure diversity, we adapt multiple diversity indices used in ecology and propose a new diversity index, the Lexicographic Counting Index, which is designed to measure

[1]Supported by Deutsche Forschungsgemeinschaft (DFG), project PACS (FL 1247/1-1, Nr. 522475669)

[2]Indeed, PB instances from Pabulib used in our experiments provide such characteristics. Nevertheless, we focus on the plain multiwinner election scenario without prices and budget in this paper.

the diversity of a committee transparently. We provide an analysis of the properties of the diversity indices, including newly introduced properties, capturing, among others, the ability of voters to easily understand why one solution is more diverse than another. With these properties, we characterize our new diversity index and are able to differentiate between any two diversity indices considered.

We also analyze the computational complexity of computing a committee for both models with most of the indices considered and scoring functions of well-known voting rules. We show that, while computing maximally diverse committees (without any satisfaction goal) is polynomial-time solvable for all indices, it is NP-hard to compute maximally diverse committees that provide some minimum satisfaction for each voter. The computational complexity of computing maximally diverse committees which provide some minimum score depends very much on the voting rule whose score we consider: E.g., this problem is polynomial-solvable for each of the indices considered when using the score of Approval Voting or Satisfaction Approval Voting, but not when using the score of Chamberlin-Courant.

Our experimental results provide insights into the influence of weakening the score or satisfaction constraint on the diversity reached. We find that the diversity of the committees determined by scoring-based approval rules can indeed be improved by using a diversity index: For example, allowing the score to be reduced by 10% of the optimal score increases the average percentage of the optimal diversity achieved by 12–19, depending on the index and score considered. However, a reduction of the score by even 50% does not lead to the optimal diversity in some cases.

In the following, proofs marked with $\star$ are deferred to the appendix.

Related Work. A significant body of literature that addresses the intersection of diversity and labelled multiwinner elections focuses on various models that include diversity constraints or quotas.

Celis et al. [10] and Bredereck et al. [9] introduce very similar models wherein candidates possess (possibly structured) labels and diversity is reached through hard distributional constraints. Their approach seeks an optimal committee that maximizes a performance score while meeting specified label occurrence requirements, such as gender quotas. In a similar vein, Ianovski [14] explore a model that accommodates "dominance constraints", requiring certain labels to occur at least as frequently as others, adding a layer of comparative label evaluation. The work by Aziz [2] provides a polynomial-time algorithm computing committees that satisfy two axioms, one ensuring the given diversity constraints are satisfied as much as possible and the other one integrating candidate excellence. Evequoz et al. [11] present an innovative election process where voters initially decide on distributional constraints for candidates' attributes before electing candidates under those constraints, demonstrating this method in a Swiss primary election study. In the domain of approval voting, Straszak et al. [33] proposed an integer linear programming (ILP) framework to address diversity constraints across categorized candidate labels, offering computational tools and real-world data applications. In a working paper, Takoulo et al. [34] approach the integration of diversity constraints within the framework of multiwinner elections, presenting a model where the committee selection prioritizes both high scores and diversity metrics. Their work extends the class of well-known committee scoring rules by tailoring them to meet specified diversity requirements and introduce new axioms for diverse committee selection under constraints.

Exploring applications beyond elections, Gawron and Faliszewski [13] utilize multiwinner voting to refine search systems such as movie recommendations, balancing similarity with diversity without relying on explicit diversity indices. The model from Relia [29] includes attributes for candidates and voters, applying hard distributional constraints and ensuring population-based representation within elected committees, enriching the voting model with demographic considerations. Lastly, Izsak et al. [16] present a framework where alternatives are classified, and inter- and intraclass relations are modeled through synergy functions, aiming to maximize both score-based and relational metrics—a concept parallel to our work, where diversity indices could be interpreted as synergy measures under score and satisfaction constraints.

Finding diverse solutions has also become important in other contexts of collective decision making.

Benabbou et al. [6] analyze diversity constraints in context of (utilitarian) public housing allocation. Aygün and Bó [1] explore diversity constraints for Brazilian college admissions through affirmative action policies, examining the strategic complexities of multidimensional privileges and proposing a fair, strategy-proof mechanism to ensure equitable student selection. Aziz and Sun [3] analyze diversity in the context of school admissions by defining a rank-based diversity concept, where maximal diversity is achieved by prioritizing student matches to seats that fulfill the institution's most crucial diversity criteria, thereby optimizing representation of key groups. Biró et al. [8] analyze the computational complexity of stable-matching-based college admissions and incorporate lower quotas for individual colleges and common quotas for groups of colleges allowing to manage collective diversity targets.

2 The Model

Let $\mathbb{N}$ and $\mathbb{N}_0$ be the natural numbers excluding and including zero, respectively, $[t]$ the set $\{1, \dots, t\}$ for any integer t, and $\mathcal{P}(X)$ the power set of any set X. We write $(x_i)_{i=a}^{b}$ short for a sequence $(x_a, x_{a+1}, \dots, x_b)$.

We consider elections of committees with approval preferences where candidates have labels, i.e., elections of the form $\mathcal{E} = (A, C, U, k, L, \lambda)$ consisting of a set A of agents, a candidate set C (e.g., projects), an approval profile $U\colon A \to \mathcal{P}(C)$, and a desired committee size $0 < k \leq |C|$. $L = \{l_1, \dots, l_m\}$ is a set of m labels (e.g., education and sport as the target of the projects) and $\lambda\colon C \to L$ assigns a label to each candidate. In addition, for $i \in [m]$ and a committee $S \subseteq C$, let

$$C_{label}(\mathcal{E}, S, i) \coloneqq \{c \in S : \lambda(c) = l_i\},$$
$$n_i(\mathcal{E}, S) \coloneqq |C_{label}(\mathcal{E}, S, i)|, \text{ and } p_i(\mathcal{E}, S) \coloneqq n_i(\mathcal{E}, S) \,/\, |S|$$

be the set, number, and proportion of candidates in S with label l_i, respectively. Furthermore, let

$$\mathrm{distr}(\mathcal{E}, S) \coloneqq (|\{i \in [m] : n_i(\mathcal{E}, S) = j\}|)_{j=0}^{|S|},$$

where the j-th entry indicates the number of labels occurring $j - 1$ times in S.

A rule $\mathcal{R}$ maps an election to at least one k-sized subset of C, i.e., $\mathcal{R}\colon \mathcal{E} \to \mathcal{P}(\{S \subseteq C : |S| = k\})$. We denote by $\mathcal{R}_{\mathrm{vld}}(\mathcal{E}) \coloneqq \{S \subseteq C : |S| = k\}$ the rule that outputs all committees of size k and by $\mathcal{R}^s(\mathcal{E}) = \arg\max_{S \in \mathcal{R}_{\mathrm{vld}}(\mathcal{E})} s(\mathcal{E}, S)$ the rule that outputs all $S \in \mathcal{R}_{\mathrm{vld}}(\mathcal{E})$ with maximal score, where s is a scoring function mapping an election and a committee to $\mathbb{N}$. We only consider scoring functions which take only A, C, U and k into account, i.e., information that is part of a "classical" election. For an $S \in \mathcal{R}_{\mathrm{vld}}(\mathcal{E})$, we measure the satisfaction of an agent as $\mathrm{sat}(\mathcal{E}, S, a) \coloneqq |S \cap U(a)|$, i.e. as the number of chosen candidates agent a approves. To measure the diversity, we look at diversity indices which assign a real number to an election $\mathcal{E}$ and committee $S \in \mathcal{R}_{\mathrm{vld}}(\mathcal{E})$. In the following, we omit arguments if they are clear from the context.

3 The Diversity Indices

In this section, we discuss the diversity indices we consider in this paper.

Example 1. As a running example, consider an election $\mathcal{E}$ with projects as candidates, $m = 3$ labels with $L = \{$🤔, 📚, 💪$\}$ (🤔 represents the label "health", 📚 "education", and 💪 "sport")[3], $k = 10$, and the following committees (for each candidate, we indicate its label):

[3]The emoji graphics are taken from `twemojis` and licensed under CC-BY 4.0: https://creativecommons.org/licenses/by/4.0/. Copyright 2019 Twitter, Inc and other contributors.

$S' \in \mathcal{R}_{\text{vld}}(\mathcal{E})$	$\text{distr}(\mathcal{E}, S')$
$S'_1 = \{$🧐, 🧐, 🧐, 📚, 📚, 📚, 💪, 💪, 💪, 💪$\}$	$(0, 0, 0, 2, 1, 0, 0, 0, 0, 0, 0)$
$S'_2 = \{$🧐, 📚, 💪, 💪, 💪, 💪, 💪, 💪, 💪, 💪$\}$	$(0, 2, 0, 0, 0, 0, 0, 0, 1, 0, 0)$
$S'_3 = \{$📚, 📚, 📚, 📚, 📚, 💪, 💪, 💪, 💪, 💪$\}$	$(1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0)$

Clearly, S'_1 appears as most diverse in the sense that it contains all three different labels and the labels appear as evenly balanced as possible. Whether S'_2 is more diverse than S'_3 is in the eye of the beholder or, put differently, depends, e.g., on whether someone finds it more important that as many labels as possible are represented or that the labels represented appear as evenly as possible: S'_2 contains three labels, but one label occurs much more frequently than the others, while S'_3 contains only two labels, but these two labels occur equally often. ◁

Indices Used in Ecology. In the field of ecology, various indices have been defined to measure the diversity of a community of species—see, e.g., [27] and [23] for an overview of diversity indices. We directly adapt the following diversity indices often used in ecology so that they receive an election $\mathcal{E} = (A, C, U, k, L, \lambda)$ as well as a committee $S \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ as inputs. For each of the following indices, a higher value indicates a higher diversity.

Richness [35] is a simple diversity index that does not take the proportions of the labels into account, but only the number of labels present:

$$Ri(\mathcal{E}, S) = |\{i \in [m] : n_i(\mathcal{E}, S) > 0\}| = \sum_{\ell=1}^{k} \text{distr}(\mathcal{E}, S)_{\ell+1} = m - \text{distr}(\mathcal{E}, S)_1.$$

The Simpson index [32][4] considers the probability that two candidates chosen independently and at random from the committee have the same label, i.e.,

$$Si(\mathcal{E}, S) = -\sum_{i \in [m]} p_i(\mathcal{E}, S)^2 = -\sum_{\ell=1}^{k} \text{distr}(\mathcal{E}, S)_{\ell+1} \cdot \left(\frac{\ell}{k}\right)^2.$$

Another popular index [24] is Shannon's entropy [31], which is derived from information theory and represents the uncertainty in predicting the label of a randomly chosen candidate from the committee:

$$Sh(\mathcal{E}, S) = -\sum_{i \in [m] : p_i(\mathcal{E}, S) > 0} p_i(\mathcal{E}, S) \cdot \log(p_i(\mathcal{E}, S)) = -\sum_{\ell=1}^{k} \text{distr}(\mathcal{E}, S)_{\ell+1} \cdot \frac{\ell}{k} \cdot \log\left(\frac{\ell}{k}\right).$$

Remark 1. The indices rank the committees from Example 1 as follows: $Ri(\mathcal{E}, S'_1) = 3 = Ri(\mathcal{E}, S'_2) > Ri(\mathcal{E}, S'_3) = 2$, $Sh(\mathcal{E}, S'_1) \approx 1.09 > Sh(\mathcal{E}, S'_3) = 0.69 > Sh(\mathcal{E}, S'_2) \approx 0.64$, and $Si(\mathcal{E}, S'_1) = -0.34 > Si(\mathcal{E}, S'_3) = -0.5 > Si(\mathcal{E}, S'_2) = -0.66$. Therefore, Ri classifies S'_3 as least diverse, while Sh and Si both classify S'_2 as the least diverse. ◁

New Index. We introduce a new diversity index, the *Lexicographic Counting Index* (LC). While the idea of lexicographic ordering has been applied in various settings, it has not been used for defining a diversity index before, to the best of our knowledge. The idea behind LC—which elevates the natural, but simple, diversity index Ri—is the following: The primary goal is to maximize the number of labels

[4] In the literature, the Simpson index is usually stated unnegated. We add the negation in order to maximize each index.

occurring at least once (like Ri does), the secondary goal is to maximize the number of labels occurring at least twice, and so on:

$$LC(\mathcal{E},S)=\sum_{i=1}^{k}(\min\{m,k\}+1)^{k+1-i}\cdot|\sigma_i(\mathcal{E},S)|\,,\text{ with }\sigma_i(\mathcal{E},S)=\{\ell\in[m]:n_\ell(\mathcal{E},S)\geq i\}\,.$$

Note that the base is $\min\{m,k\}+1$, as a committee consists of k candidates and each candidate introduces at most one new label, i.e. $\sigma_i(\mathcal{E},S)\leq\min\{m,k\}$ for all $i\in[k]$. In Appendix A.1, we evaluate LC with respect to properties adapted from the literature: These results provide some arguments why the new index can be called a diversity index.

Remark 2. LC classifies S_1' as the most and S_3' as the least diverse committee: $LC(\mathcal{E},S_1')=4^{10}\cdot3+4^9\cdot3+4^8\cdot3+4^7=4145152>LC(\mathcal{E},S_2')=4^{10}\cdot3+4^9+4^8+4^7+4^6+4^5+4^4+4^3=3495232>LC(\mathcal{E},S_3')=4^{10}\cdot2+4^9\cdot2+4^8\cdot2+4^7\cdot2+4^6\cdot2=2793472$. ◁

4 Which diversity indices to use?

Next, we want to distinguish formally between the aforementioned indices by defining properties and testing the indices against them. Note that not all properties that we define should necessarily be satisfied by a diversity index: The properties rather draw attention to differences between the indices, which should be taken into account when picking a diversity index to be used. This is of interest when electing committees consisting of at least six candidates, because Sh, Si, and LC behave the same for small committees when deciding which committee is more diverse:

Observation 1 (⋆). *For all elections $\mathcal{E}$ with $k\leq5$, it holds that $\forall r,r'\in\{Sh,Si,LC\},\diamond\in\{<,>,=\},S_1,S_2\in\mathcal{R}_{\text{vld}}(\mathcal{E}):r(\mathcal{E},S_1)\diamond r(\mathcal{E},S_2)\Leftrightarrow r'(\mathcal{E},S_1)\diamond r'(\mathcal{E},S_2)$.*
In addition, for all elections $\mathcal{E}$ with $k\leq7$, it holds that $\forall S_1,S_2\in\mathcal{R}_{\text{vld}}(\mathcal{E}),\diamond\in\{<,>,=\}:Sh(\mathcal{E},S_1)\diamond Sh(\mathcal{E},S_2)\Leftrightarrow LC(\mathcal{E},S_1)\diamond LC(\mathcal{E},S_2)$.

However, the indices can behave differently for larger k. One reason for this is that only Ri and LC consider the number of labels present to be more important than the evenness of the distribution of the labels present, while Si and Sh do not—this can be seen, e.g., in Example 1. Thus, the first question that one has to answer is whether having as many labels in the committees as possible has the highest priority—this could be the case when electing a team working on an interdisciplinary project, where having an expert from a discipline that is not covered otherwise is very valuable. We express this through the following property:

Property 1 (Present Label Maximization)**.** A diversity index D satisfies *Present Label Maximization* if, for all elections $\mathcal{E}$ and $S_1,S_2\in\mathcal{R}_{\text{vld}}(\mathcal{E})$ for which $\text{distr}(\mathcal{E},S_1)_1<\text{distr}(\mathcal{E},S_2)_1$ (i.e., S_1 contains more different labels than S_2), it holds that $D(S_1)>D(S_2)$.

Observation 2 (⋆). *Ri and LC satisfy Present Label Maximization, Si and Sh do not.*

Another property which is quite natural and which should be satisfied by any diversity index is that increasing the occurrences of a label by decreasing those of a more frequent label by at least two leads to a higher (*Occurrence Balancing*) or at least the same (*Weak Occurrence Balancing*) diversity:

Property 2 (Occurrence Balancing)**.** A diversity index D satisfies (*Weak*) *Occurrence Balancing* if for all elections $\mathcal{E}$ and $S'\in\mathcal{R}_{\text{vld}}(\mathcal{E})$ for which there are $i,j\in[m]$ with $n_i(\mathcal{E},S')+2\leq n_j(\mathcal{E},S')$ and $n_i(\mathcal{E},S')<|C_{label}(\mathcal{E},C,i)|$, it holds that $\forall c_i\in C_{label}(\mathcal{E},C,i)\setminus S',c_j\in C_{label}(\mathcal{E},S',j):D(S')<D(S'')$ ($D(S')\leq D(S'')$) with $S''=S'\setminus\{c_j\}\cup\{c_i\}$.

The following result not only separates Ri from the other three indices, but is also the basis for showing in Section 5.1 that finding a committee with the highest diversity—according to one of the diversity indices considered—is in $\mathbf{P}$.

Observation 3 (⋆). *Of the diversity indices considered, all satisfy Weak Occurrence Balancing, but only Si, Sh, and LC satisfy Occurrence Balancing.*

The properties introduced so far allow us to differentiate between any pair of diversity indices considered apart from Si and Sh, which both take the evenness of the distribution into account. However, these two indices differ in whether it is better to balance the occurrences of two labels when these labels are relatively rare or relatively dominant. For this, we first define which pairs of labels whose occurrences differ by d are *balancable*: $\text{balancable}(\mathcal{E}, S, d)$ returns, for a given election $\mathcal{E}$ and a committee $S \in \mathcal{R}_{\text{vld}}(\mathcal{E})$, all pairs (i, j) so that $n_i(\mathcal{E}, S) + d = n_j(\mathcal{E}, S)$ and $n_i(\mathcal{E}, S) + \lfloor \frac{d}{2} \rfloor \leq |C_{label}(\mathcal{E}, C, i)|$ (i.e. the frequency of l_i could be increased by $\lfloor \frac{d}{2} \rfloor$ and the first label of the pair occurs d fewer times). The function $\text{balance}(\mathcal{E}, S, d, (i, j))$ actually balances such labels by taking, in addition to $\mathcal{E}$ and S, a label pair $(i, j) \in \text{balancable}(\mathcal{E}, S, d)$ as an argument and returning a committee $S' \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ so that $n_i(\mathcal{E}, S') = n_i(\mathcal{E}, S) + \lfloor \frac{d}{2} \rfloor, n_j(\mathcal{E}, S') = n_j(\mathcal{E}, S) - \lfloor \frac{d}{2} \rfloor$ and $\forall e \in [m] \setminus \{i, j\} : n_e(\mathcal{E}, S') = n_e(\mathcal{E}, S)$, i.e. only the number of l_i and l_j are balanced. Based on this, we can define the following property:

Property 3 (Prioritization of Rare Label Balancing)**.** A diversity index D satisfies *Prioritization of Rare Label Balancing* if, for all elections $\mathcal{E}$, $S \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ and $d \geq 2$ for which there are $(i, j), (k, l) \in \text{balancable}(\mathcal{E}, S, d)$ with $n_i(\mathcal{E}, S) < n_k(\mathcal{E}, S)$, it holds that $D\big(\mathcal{E}, S_{(i,j)}\big) > D\big(\mathcal{E}, S_{(k,l)}\big)$ with $S_{(i,j)} \coloneqq \text{balance}(\mathcal{E}, S, d, (i, j))$ and $S_{(k,l)} \coloneqq \text{balance}(\mathcal{E}, S, d, (k, l))$.

Observation 4 (⋆). *Sh and LC satisfy Prioritization of Rare Label Balancing, Si and Ri do not.*

In some sense, this makes Sh more similar to LC than Si, as Sh and LC both prefer to increase the frequency of the label that is rarest among the four labels. In contrast, Si rates both options as equally good and thus does not differentiate whether you decrease the frequency of the most dominant label or increase the frequency of the rarest label among the four labels at hand.

Next, we want to look at another property that separates Si and Sh from LC and that takes into account that the index is to be used in an election: One important goal is to ensure that voters (or other stakeholders) are able to understand why a committee has been chosen over a different one, which is likely to promote acceptance of the result or at least a more informed debate about it (e.g., when maximizing diversity is incorporated into elections). Hence, we want to focus on the *obviousness* of the diversity indices next or, in other words, the effort required to decide which of two given committees is more diverse according to the used diversity index. The only information necessary for this is how many labels occur how often, which is provided by the distr vectors. Consider, for example, S'_2 and S'_3 from Example 1 and their distr vectors and try to answer which of the committees is more diverse according to one of the indices considered. When considering this question for the Shannon entropy, for instance, the question probably cannot be answered without calculating the diversities with Sh's formula. This problem does not necessarily apply to each of the presented diversity indices:

Let D be one of the considered diversity indices and $\mathcal{E}$ some election such that there are two differently diverse committees $S_1, S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E})$. When comparing $\text{distr}(\mathcal{E}, S_1)$ and $\text{distr}(\mathcal{E}, S_2)$ to determine which one is more diverse, any $i \in [k+1]$ with $\text{distr}(\mathcal{E}, S_1)_i = \text{distr}(\mathcal{E}, S_2)_i$ is irrelevant. Let $I_R(\mathcal{E}, S_1, S_2) = \{i \in [k+1] : \text{distr}(\mathcal{E}, S_1)_i \neq \text{distr}(\mathcal{E}, S_2)_i\}$ be the set of all indices that hence matter. Note that $I_R(\mathcal{E}, S_1, S_2)$ is non-empty; otherwise, the two committees would have the same diversity (regardless of the index used). Based on this, we define $\text{rdistr}(\mathcal{E}, S_1, S_2)$ as the distr vector of S_1 at the indices in $I_R(\mathcal{E}, S_1, S_2)$, with ρ as the vector of the elements of $I_R(\mathcal{E}, S_1, S_2)$ in ascending order:

$$\text{rdistr}(\mathcal{E}, S_1, S_2) \coloneqq (\text{distr}_{\rho_i}(\mathcal{E}, S_1))_{i=1}^{|I_R(\mathcal{E}, S_1, S_2)|}$$

Remark 3. For the committees S'_2 and S'_3 in Example 1, we have $I_R(\mathcal{E}, S'_2, S'_3) = I_R(\mathcal{E}, S'_3, S'_2) = \{1, 2, 6, 9\}$, $\text{rdistr}(\mathcal{E}, S'_2, S'_3) = (0, 2, 0, 1)$, and $\text{rdistr}(\mathcal{E}, S'_3, S'_2) = (1, 0, 2, 0)$. ◁

Based on this, the following property can be defined:

Property 4 (Obviousness)**.** A diversity index D is *obvious* if there is a function $f\colon \mathbb{N} \to \mathbb{N}$ such that for every election $\mathcal{E}$ and $S_1, S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ for which $D(\mathcal{E}, S_1) \neq D(\mathcal{E}, S_2)$ and $l = |I_R(\mathcal{E}, S_1, S_2)| > 0$, it holds that $D(\mathcal{E}, S_1) > D(\mathcal{E}, S_2)$ if and only if $\text{rdistr}(\mathcal{E}, S_1, S_2)_{f(l)} < \text{rdistr}(\mathcal{E}, S_2, S_1)_{f(l)}$.

Intuitively, f maps to the index of rdistr where we can see which of two committees is more diverse. Natural examples are $f(l) = 1$ for all $l \in \mathbb{N}$, i.e., mapping to the first index, and $f(l) = l$ for all $l \in \mathbb{N}$, i.e., mapping to the last index. For the indices we consider, we have the following which tells us that it should be kept in mind that understanding why a committee is chosen over another when using Si or Sh can be challenging:

Observation 5 (⋆)**.** *Ri and LC are obvious with $f(l) = 1$ for all $l \in \mathbb{N}$. Sh and Si are not obvious.*

Remark 4. Using Observation 5, we can directly see from the rdistr vectors given in Remark 3 that S_2' is more diverse than S_3' according to LC and Ri. ◁

Remark 5. $f(l) = l$ holds, e.g., for the Berger-Parker index [7], which is defined as $-\max_{i\in[m]} p_i(\mathcal{E}, S)$—in the literature, it is usually defined without negation—and hence measures the dominance of the most frequent label. We do not investigate this index further to have a clear focus.

Characterizing LC. As LC is the only diversity index considered that fulfills all properties introduced so far, the question arises whether they characterize the behavior of LC when comparing two committees $S_1, S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E})$. This is not the case:

Example 2. Consider an election $\mathcal{E}$ with $|C| = 12$, labels l_1, l_2, l_3, l_4 with $|C_{label}(\mathcal{E}, C, 1)| = 2$, $|C_{label}(\mathcal{E}, C, 2)| = 3$, $|C_{label}(\mathcal{E}, C, 3)| = 3$, $|C_{label}(\mathcal{E}, C, 4)| = 4$, and $k = 10$, which results in the following possible distr vectors (we only indicate the first five elements as the rest are zeros):

$$\text{distr}(\mathcal{E}, S_1) = (1, 0, 0, 2, 1)\,, \text{distr}(\mathcal{E}, S_2) = (0, 1, 1, 1, 1)\,, \text{distr}(\mathcal{E}, S_3) = (0, 1, 0, 3, 0)\,,$$
$$\text{distr}(\mathcal{E}, S_4) = (0, 0, 3, 0, 1)\,, \text{distr}(\mathcal{E}, S_5) = (0, 0, 2, 2, 0)\,.$$

It holds that $LC(\mathcal{E}, S_1) < LC(\mathcal{E}, S_2) < LC(\mathcal{E}, S_3) < LC(\mathcal{E}, S_4) < LC(\mathcal{E}, S_5)$. However, a diversity index D with $D(\mathcal{E}, S_1) < D(\mathcal{E}, S_2) < D(\mathcal{E}, S_3) = D(\mathcal{E}, S_4) < D(\mathcal{E}, S_5)$ also satisfies the properties.◁

We therefore introduce another property which makes it impossible to rank S_3 and S_4 as equally diverse in Example 2 and which expresses that an index is quite fine-grained:

Property 5 (Distribution Equivalence)**.** A diversity index D satisfies *Distribution Equivalence* if for all elections $\mathcal{E}$ and $S_1, S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ it holds that $D(\mathcal{E}, S_1) = D(\mathcal{E}, S_2) \Leftrightarrow \text{distr}(\mathcal{E}, S_1) = \text{distr}(\mathcal{E}, S_2)$.

Observation 6 (⋆)**.** *Of the diversity indices considered, only LC satisfies Distribution Equivalence.*

This observation tells us that, if the distr vectors of two committees are different, LC will always rank one committee as more diverse, which ensures that no additional tie-breaking (which can bear the potential for conflict) between committees with different distr vectors (which could otherwise be categorized as equally diverse) is necessary. This is not the case, e.g., for Ri—which is the only other diversity index considered that satisfies *Obviousness*. Thus, LC is the only diversity index that satisfies all the properties discussed so far, three of which suffice to characterize LC:

Theorem 1 (⋆)**.** *LC is characterized by the properties Distribution Equivalence, Obviousness, and Present Label Maximization.*

5 Incorporating Diversity Indices into Elections

In this section, we propose two ways to incorporate diversity (indices) into elections. One way is to maximize the diversity given a minimal satisfaction for each agent:

Definition 1 (MAX-D-DSAT)**.** Given an election $\mathcal{E} = (A, C, U, k, L, \lambda)$ and a function $h : A \to \mathbb{N}_0$, find a committee with maximal diversity with respect to the diversity index $\mathcal{D}$ among all committees $S \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ for which $\forall a \in A : \text{sat}(\mathcal{E}, S, a) \geq h(a)$.

On the one hand, this provides a certain amount of freedom in the search for a diverse committee, but on the other hand, it gives the voters the certainty that their satisfaction cannot be worsened arbitrarily. One possibility for defining $h(a)$ is to compute a committee S with a well-known rule and to define $h(a)$ as the satisfaction of agent a with S minus one (which we will consider in our numerical experiments) or to ensure the same minimum satisfaction for each agent.

A different way to incorporate diversity is to maximize the diversity given a scoring function (e.g., of a scoring-based voting rule) and a minimum committee score:

Definition 2 (MAX-(D, s)-DSCR)**.** Given an election $\mathcal{E}$ and a bound $\beta \in \mathbb{N}_0$, find a committee with maximal diversity with respect to diversity index D among all committees $S \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ for which $s(\mathcal{E}, S) \geq \beta$, where s is a scoring function.

We will show results for the following, well-known scoring-based approval voting rules in the following sections: Multi-Winner Approval Voting (AV), Satisfaction Approval Voting (SAV), Proportional Approval Voting (PAV), and Approval Chamberlin-Courant (CC) (see, e.g., [19] for a definition of these rules). We will refer to the score of the scoring-based voting rule $\mathcal{R}$ as $\text{score}_{\mathcal{R}}$.

In the following, we refer to the decision variant of MAX-D-DSAT and MAX-(D, s)-DSCR, in which the goal is to find a committee with a diversity which is at least a given value δ, as D-DSAT and (D, s)-DSCR, respectively.

5.1 Complexity Results

First, we consider the computational complexity of finding a "diversity optimal" committee without additional constraints. For this, we can exploit the fact that each index satisfies *Weak Occurrence Balancing*. Therefore, the highest diversity is reached when starting with an empty committee and iteratively adding a candidate with a label that, among the labels with unselected candidates, occurs least often in the current committee:

Observation 7 (⋆)**.** *Given an election $\mathcal{E}$ and one of the diversity indices considered, choosing a committee $S \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ with the highest diversity is polynomial-time solvable.*

Yet, D-DSAT is **NP**-hard for these indices. The same holds for (D, s)-DSCR if finding winning committees is **NP**-hard for $\mathcal{R}^s$. We show this result for a broader class of diversity indices satisfying the following, clearly desirable property:

Property 6 (Uniqueness Optimality)**.** A diversity index D satisfies *Uniqueness Optimality* if for all elections with $m \geq k$, it holds that $S \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ has optimal diversity if and only if no two candidates in S have the same label.

Clearly, each of LC, Si, and Sh satisfy this property because they satisfy *Occurrence Balancing*, and Ri satisfies it because it is defined as $m - \text{distr}(\mathcal{E}, S)_1$ and $\text{distr}(\mathcal{E}, S)_1$ gets larger if labels occur more than once. For such indices, we have the following:

Observation 8. *If D is a diversity index that satisfies Uniqueness Optimality, (1) (D, s)-DSCR is **NP**-hard if the decision problem of $\mathcal{R}^s$ is **NP**-hard and (2) D-DSAT is **NP**-hard.*

Proof. (1) The reduction from the **NP**-hard decision problem of $\mathcal{R}^s$, i.e., where given an election $\mathcal{E} = (A, C, U, k)$, the task is to decide whether there is $S \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ with $s(\mathcal{E}, S) \geq s^*$, works as

follows: Choose the election $\mathcal{E}' = (A, C, U, k, L', \lambda')$ with $L' = \{l_c : c \in C\}$ as the possible labels and $\lambda'(c) = l_c$ (i.e., each candidate has a different label). Set $\delta = D(\mathcal{E}, S^*)$, where S^* is a committee for $\mathcal{E}$ in which k labels occur exactly once, and $\beta = s^*$. Thus, each committee of size k for $\mathcal{E}'$ has the diversity $D(\mathcal{E}, S^*)$. Hence, (D, s)-DSCR has a solution if and only if there is a committee of score at least s^*. (2) The reduction from the problem of finding a committee in which each agent has a satisfaction of at least one, which is **NP**-hard [28], works analogously by giving each candidate a different label, setting δ to $D(\mathcal{E}, S^*)$ with S^* being a committee with k labels, and $h(a) = 1$ for all agents $a \in A$. □

As the decision problems of CC and PAV are **NP**-hard [28, 4], we have that $(D, \mathrm{score}_{\mathrm{CC}})$-DSCR and $(D, \mathrm{score}_{\mathrm{PAV}})$-DSCR are **NP**-hard, where D is one of the indices considered.

In the remainder of this section, we want to focus on separable scoring functions:

Definition 3. A scoring function s is *separable* if there is a polynomial-time computable function w mapping an election and a candidate to $\mathbb{N}$ such that for every election $\mathcal{E}$ and for every $S \subseteq C$, we have that $s(\mathcal{E}, S) = \sum_{c \in S} w(\mathcal{E}, c)$.

Clearly, a committee of $\mathcal{R}^s$ can be computed in polynomial time if s is separable. The scoring function of AV is such a separable function with $w(\mathcal{E}, c) = |\{a \in A : c \in U(a)\}| \leq |A|$, and the scoring function of SAV can be transformed into a separable function with $w(\mathcal{E}, c) = \sum_{a \in A : c \in U(a)} (\ell / |U(a)|)$, where ℓ is the least common multiple of $\bigcup_{a \in A} \{|U(a)|\}$, which can be computed in polynomial time using binary representation. For separable functions, the following holds:

Theorem 2 (⋆)**.** MAX-(D, s)-DSCR *is in* **P** *if s is a separable function and $D \in \{Ri, LC\}$.*

Therefore, for $D \in \{Ri, LC\}$, it holds that MAX-$(D, \mathrm{score}_{\mathrm{AV}})$-DSCR and MAX-$(D, \mathrm{score}_{\mathrm{SAV}})$-DSCR are in **P**. The algorithm for Theorem 2 utilizes the lexicographic nature of LC. Next we show tractability results for a class of diversity indices which we call label-wise diminishing, which includes Si and Sh. An index D is called *label-wise diminishing* if D can be expressed as $\sum_{l \in [m]} \sum_{i=1}^{n_l} t(i)$ and, for all $i \in [k]$, it holds that $0 < t(i)$ and $t(i)$ can be computed in time polynomial in the input size, and t is strictly monotonically decreasing. We have the following.

Theorem 3. MAX-(D, s)-DSCR *is in* **P** *if*

- *s is a separable function and there is an $\alpha \in \mathbb{N}_0$ polynomial in the input size so that, for each $c \in C$, $w(\mathcal{E}, c) \leq \alpha$,*
- *D is label-wise diminishing.*

Proof. Given an instance I_D of MAX-(D, s)-DSCR with n candidates, we construct an instance I_K of the 0-1 Knapsack problem in polynomial time. We assume that $\beta \leq k \cdot \alpha$ (β being the lower bound for the score, see Definition 2), since there is no solution for I_D otherwise. For each candidate c_i, we add an item x_i with the weight $w_K(x_i) \coloneqq n\alpha + 1 - w(c_i)$ and the value $v(x_i) = t(\pi(c_i)) + \eta$, where $\eta \coloneqq k \cdot t(1) + 1$ and π outputs c_i's position in a descending ordering of the candidates with the same label as c_i based on w. The knapsack's bound is $B \coloneqq k(n\alpha + 1) - \beta$. Let, for a solution X of I_K, $S(X) = \{c_i \mid x_i \in X\}$, $v(X)$ and $w_K(X)$ the value and weight of X, and, for a solution S of I_D, $X(S) = \{x_i : c_i \in S\}$. We claim that I_K has a solution X with value at least $k \cdot \eta$ if and only if $S(X)$ is a solution to I_D and that I_D is infeasible otherwise:

Let X be a solution to I_K with value at least $k \cdot \eta$. It follows that $|X| \geq k$. Assume that $|X| > k$, then $w_K(X) \geq (k+1)(n\alpha+1) - w(S(X)) \geq (k+1)(n\alpha+1) - n\alpha = k(n\alpha+1)+1 > B$, a contradiction. Thus, $|X| = k$ and $w_K(X) \leq B \Leftrightarrow k(n\alpha + 1) - w(S(X)) \leq k(n\alpha + 1) - \beta \Leftrightarrow \beta \leq w(S(X))$. Thus, $S(X)$ is of size exactly k and respects the score constraint. If X is a solution to I_K with value less than $k \cdot \eta$, then $|X| < k$. Since X is maximal, there is no size-k solution respecting the capacity

constraints and hence no solution to I_D that respects the score constraint. Analogously, I_K has no solution with value at least $k \cdot \eta$ if I_D has none.

Finally, we show that, for an optimal solution X_K^* of I_K with value at least $k \cdot \eta$, $S(X_K^*)$ is an optimal solution of I_D. Note that, if, for a label l_j, n_j many items x_i with $\lambda(c_i) = l_j$ are chosen in X_K^*, the n_j items which come first in the order π are always chosen and thus $v(X_K^*) = D(\mathcal{E}, S(X_K^*)) + k \cdot \eta$. Next, assume that I_D has an optimal solution $S_D^* \neq S(X_K^*)$ with $D(S_D^*) > D(S(X_K^*))$. Consider the committee S^* with $n_l(\mathcal{E}, S^*) = n_l(\mathcal{E}, S_D^*)$ for $l \in [m]$ and $c \in S^* \Leftrightarrow \pi(c) \leq n_j(\mathcal{E}, S_D^*)$ with $l_j = \lambda(c)$. Thus, $D(\mathcal{E}, S^*) = D(\mathcal{E}, S_D^*) = v(X(S^*)) - k \cdot \eta > D(\mathcal{E}, S(X_K^*)) = v(X_K^*) - k \cdot \eta$, a contradiction.

Thus, the problem can be solved, e.g., by the solving the 0-1 Knapsack instance with dynamic programming in $\mathcal{O}(nB) = \mathcal{O}(n\,(k(n\alpha + 1) - \beta))$ time. □

Si is label-wise diminishing with $t(i) = 2k + 1 - 2i$, and Sh is label-wise diminishing with $t(i) = -i\log(i) + (i-1)\log(i-1) + \log(k) + 2$ where $t(1) = \log(k) + 2$. Thus, we have:

Corollary 1 (⋆). MAX-$(Si, \mathrm{score}_{\mathrm{AV}})$-DSCR *is in* $\mathbf{P}$ *and, if considering a computational model in which the logarithm of natural numbers and addition and multiplication including a logarithm can be computed in polynomial time,* MAX-$(Sh, \mathrm{score}_{\mathrm{AV}})$-DSCR *is in* $\mathbf{P}$.

However, we cannot apply Theorem 3 for SAV with the previously mentioned approach to transform SAV's score into a separable function, as it leads to weights that could not be bounded by a value polynomial in the input size. Therefore, we show the following, which is also applicable to SAV:

Theorem 4 (⋆). (D, s)-DSCR *is in* $\mathbf{P}$ *if*

- s *is a separable function and*
- D *is label-wise diminishing and there is a* $\zeta \in \mathbb{N}$ *polynomial in the input size such that* $0 < t(i) \leq \zeta$ *for all* $i \in [k]$.

The proof of Theorem 4 is similar to the proof of Theorem 3 in terms of constructing an instance of the 0-1 knapsack problem, but the diversity constraint is expressed with the help of the weights and the score constraint with the help of the values of the items. While Sh does not fulfil the imposed conditions of Theorem 4—leaving the question open whether $(Sh, \mathrm{score}_{\mathrm{SAV}})$-DSCR is in $\mathbf{P}$—the previously mentioned choice of $t(i)$ for Si fulfills them. Therefore:

Corollary 2. $(Si, \mathrm{score}_{\mathrm{SAV}})$-DSCR *is in* $\mathbf{P}$.

5.2 Experiments

To evaluate the problems, we use datasets with approval preferences from Pabulib [12]—a collection of participatory budgeting data, a scenario in which incorporating diversity may be desirable—, in which categories (e.g., urban greenery) and/or targets (e.g., adults) are assigned to the candidates: For each such instance, we create up to three new instances of our model by assigning to a candidate as the label (1) the categories, (2) the targets, or (3) the union of the categories and targets (e.g., {urban greenery, adults} as a label and different sets forming different labels). We also transformed two datasets [21, 22] with approval preferences from PrefLib [26] about the French presidential election in 2002, consisting of seven instances overall, by assigning the combination of gender and political leaning as the label to each candidate. The dimensions of the experimental data can be seen in Fig. 1. We removed instances with $|C| = m$ because each committee leads to the optimal diversity for them, as each diversity index considered satisfies *Uniqueness Optimality*. In the following, we show results for $k = 10$. For this, we discarded instances with $|C| \leq 10$, which results in 687 instances. We conducted the same experiments for $k \in \{6, 8\}$ as well, the results of which are very similar and shown in Section D.

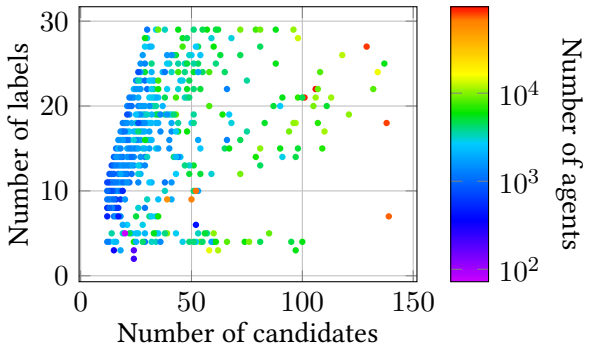


Figure 1: The dimensions of the experimental data, where the color of each point represents the average number of agents of all instances with the given number of labels and candidates.

Experimental Setup. To investigate the influence of weakening the score constraint on the diversity reached (i.e. MAX-(D, s)-DSCR), we consider score_{AV} and $\text{score}_{\text{SAV}}$ (for which solving MAX-(D, s)-DSCR in $\mathbf{P}$). For a given diversity index, let $\mathcal{R}^p_{\text{scr}}$ be the rule returning the committees with the highest diversity among the committees reaching at least $p\%$ of the highest value of $\text{score}_{\mathcal{R}}$.

To examine the influence of the satisfaction constraints (i.e. MAX-D-DSAT), we additionally consider CC, PAV, the Method of Equal Shares (Rule X), and Phragmén's sequential rule (seq-Phragmén) (see [19] for definitions) in Section D due to lack of space[5]. For each rule $\mathcal{R}$ considered, we first compute one committee S using the Python library *abcvoting* [20] with default parameters and refer to the rule returning this committee as $\mathcal{R}$. Based on the satisfactions of the agents with S, we look at the rule that returns the committees with the highest diversity reachable when the satisfaction of each agent can be decreased by at most one, which we denote as $\mathcal{R}^{-1}_{\text{sat}}$.

We also compute all winning committees for these rules with *abcvoting* to investigate whether the diversity index could serve as a tiebreaker between them.

Experimental Results. Overall, the results—many of which can be seen from Fig. 2—are very similar for AV and SAV and the following observations hold for both these rules: They achieve (without score or satisfaction constraints) around 80% of the optimal diversity when measured with $D \in \{Ri, LC, Sh\}$ and around 70% when measured with Si. Thus, the diversity of the committees can indeed be improved. In addition, for each diversity index, there are instances for which even allowing a score reduction of 50% does not lead to the optimal diversity. Choosing the winning committee of AV or SAV with the highest diversity rarely makes a difference. The most fundamental reason for this is that AV has multiple winning committees for only around 5% of the instances and SAV even more rarely.[6]

We also compare the benefits of using $\mathcal{R}^{-1}_{\text{sat}}$ or $\mathcal{R}^{90}_{\text{scr}}$ instead of $\mathcal{R}$: The experiments show that, on average, a higher proportion of the optimal diversity is reached when using $\mathcal{R}^{90}_{\text{scr}}$ than when using $\mathcal{R}^{-1}_{\text{sat}}$ (with only a few exceptions for $k = 6$ where these two approaches achieve very similar results overall): On average, $\mathcal{R}^{90}_{\text{scr}}$ reaches $4 - 7\%$ of the optimal diversity more than $\mathcal{R}^{-1}_{\text{sat}}$ (for $k = 10$). In addition, $\mathcal{R}^{90}_{\text{scr}}$ can lead to a noticeable change compared to $\mathcal{R}$: The percentage of the optimal diversity achieved on average increases by 12–19. It is also interesting that the gain in diversity is larger overall for $\mathcal{R}^{90}_{\text{scr}}$ compared to $\mathcal{R}$ than for $\mathcal{R}^{80}_{\text{scr}}$ compared to $\mathcal{R}^{90}_{\text{scr}}$. The same holds true for the gain from $\mathcal{R}^{90}_{\text{scr}}$ to $\mathcal{R}^{80}_{\text{scr}}$ compared to that from $\mathcal{R}^{80}_{\text{scr}}$ to $\mathcal{R}^{70}_{\text{scr}}$ (which can be seen in Section D), which suggests that there are diminishing returns when weakening the score constraint. When comparing the four different diversity indices, it seems most challenging to achieve the optimal diversity when using Si for each rule and approach visualized in Fig. 2.

[5] seq-Phragmén, Rule X, and PAV perform very similarly overall to AV and SAV with regard to the diversity reached (with or without satisfaction constraints), while CC reaches (slightly) higher diversities than the other rules on average.

[6] Using diversity as a tiebreaker leads to the greatest improved among all considered rules for CC (see Section D) which has multiple winning committees for 209 instances.

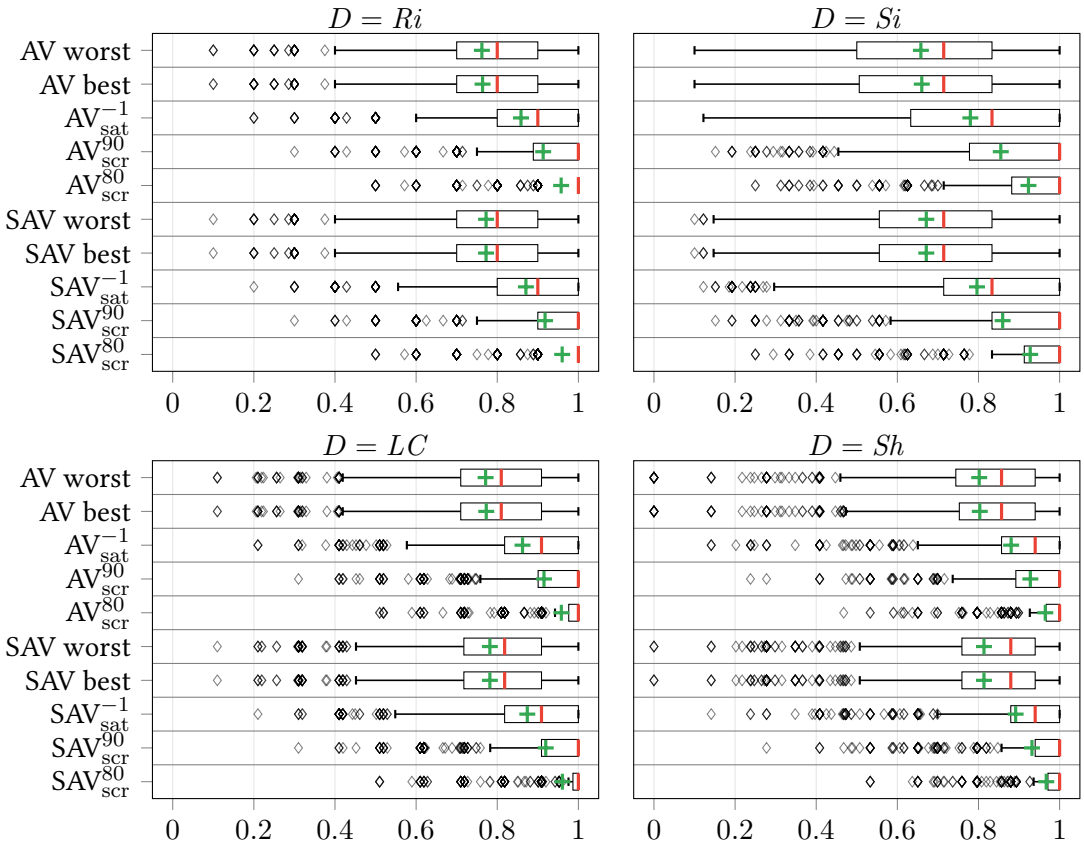


Figure 2: The proportion of the optimal diversity reached on the experimental data when using the specified diversity index D. "$\mathcal{R}$ best" ("$\mathcal{R}$ worst") refers to the rule choosing the committees with the highest (lowest) diversity among the winning committees of $\mathcal{R}$. The red line indicates the median, the green cross the mean.

6 Epilogue

We adapted several diversity indices used in ecology to the context of committee elections and introduced a new diversity index. We also introduced properties of diversity indices which allow us to differentiate between any pair of indices we consider and tested the indices against them. Finally, we characterized the new index via three of our properties. The underlying model assumes that each candidate has one label: While this allows to define a label as a set of "sub-labels" (e.g. $\{$urban greenery, adults$\}$ as one label), all the indices we consider do not take the (dis)similarity or the importance of labels into account. Further research could investigate diversity indices that incorporate such distances or different label priorities, which also requires thinking about how such distances and priorities are determined.

From an algorithmic point of view, we proved that (D, s)-DSCR is in $\mathbf{P}$ in some cases if s is a separable scoring function, despite, e.g., Si having a quadratic objective function. This includes the score of AV and SAV for each diversity index considered apart from Sh in case of SAV—we left open whether $(Sh, \text{score}_{\text{SAV}})$-DSCR is polynomial-time solvable. However, there are other s for which we prove that (D, s)-DSCR is $\mathbf{NP}$-hard, which is also the case for D-DSAT. Further work may study parameterized complexity or approximation algorithms for these problems.

Our experiments revealed interesting trade-offs between satisfaction/scoring guarantees and diversity, showing, among other results, that the diversity of committees can indeed be improved. It would also be interesting to investigate how much the diversity indices differ on real world data or to evaluate past elections regarding their scoring-diversity performance. This could be particularly interesting in the context of participatory budgeting, which calls for an extension of our model in which the costs of the projects and the respective budget becomes the third objective (next to voter satisfaction and diversity).

References

[1] Orhan Aygün and Inácio Bó. College admission with multidimensional privileges: The brazilian affirmative action case. *American Economic Journal: Microeconomics*, 13(3):1–28, 2021.

[2] Haris Aziz. A rule for committee selection with soft diversity constraints. *Group Decision and Negotiation*, 28(6):1193–1200, 2019.

[3] Haris Aziz and Zhaohong Sun. Multi-rank smart reserves: A general framework for selection and matching diversity goals. *Artificial Intelligence*, 339:104274, 2025.

[4] Haris Aziz, Serge Gaspers, Joachim Gudmundsson, Simon Mackenzie, Nicholas Mattei, and Toby Walsh. Computational aspects of multi-winner approval voting. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'15)*, pages 107–115. ACM, 2015.

[5] AJ Baczkowski, DN Joanes, and GM Shamia. Properties of a generalized diversity index. *Journal of Theoretical Biology*, 188(2):207–213, 1997.

[6] Nawal Benabbou, Mithun Chakraborty, Xuan-Vinh Ho, Jakub Sliwinski, and Yair Zick. Diversity constraints in public housing allocation. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '18)*, pages 973–981. IFAAMAS / ACM, 2018.

[7] Wolfgang H. Berger and Frances L. Parker. Diversity of planktonic foraminifera in deep-sea sediments. *Science*, 168(3937):1345–1347, 1970.

[8] Péter Biró, Tamás Fleiner, Robert W. Irving, and David F. Manlove. The college admissions problem with lower and common quotas. *Theoretical Computer Science*, 411(34):3136–3153, 2010.

[9] Robert Bredereck, Piotr Faliszewski, Ayumi Igarashi, Martin Lackner, and Piotr Skowron. Multiwinner elections with diversity constraints. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, pages 933–940. AAAI Press, 2018.

[10] L. Elisa Celis, Lingxiao Huang, and Nisheeth K. Vishnoi. Multiwinner voting with fairness constraints. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*, pages 144–151. International Joint Conferences on Artificial Intelligence Organization, 2018.

[11] Florian Evequoz, Johan Rochel, Vijay Keswani, and L Elisa Celis. Diverse representation via computational participatory elections-lessons from a case study. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO'22)*, pages 1–11. ACM, 2022.

[12] Piotr Faliszewski, Jarosław Flis, Dominik Peters, Grzegorz Pierczyński, Piotr Skowron, Dariusz Stolicki, Stanisław Szufa, and Nimrod Talmon. Participatory budgeting: Data, tools and analysis. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI-2023, pages 2667–2674. International Joint Conferences on Artificial Intelligence Organization, 8 2023. Main Track.

[13] Grzegorz Gawron and Piotr Faliszewski. Using multiwinner voting to search for movies. In *Proceedings of the 19th European Conference on Multi-Agent Systems (EUMAS'22)*, volume 13442 of *LNCS*, pages 134–151. Springer, 2022.

[14] Egor Ianovski. Electing a committee with dominance constraints. *Annals of Operations Research*, 318(2):985–1000, 2022.

[15] János Izsák. Sensitivity profiles of diversity indices. *Biometrical journal*, 38(8):921–930, 1996.

[16] Rani Izsak, Nimrod Talmon, and Gerhard Woeginger. Committee selection with intraclass and interclass synergies. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, pages 1071–1078. AAAI Press, 2018.

[17] János Izsák and László Papp. A link between ecological diversity indices and measures of biodiversity. *Ecological Modelling*, 130(1):151–156, 2000.

[18] Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

[19] Martin Lackner and Piotr Skowron. *Multi-Winner Voting with Approval Preferences*. Springer Briefs in Intelligent Systems. Springer, 2023.

[20] Martin Lackner, Peter Regner, and Benjamin Krenn. abcvoting: A Python package for approval-based multi-winner voting rules. *Journal of Open Source Software*, 8(81):4880, 2023.
[21] Jean-François Laslier and Karine Van der Straeten. A live experiment on approval voting. *Experimental Economics*, 11(1):97–105, 2008.
[22] Jean-François Laslier and Karine Van der Straeten. Une expérience de vote par assentiment lors de l'élection présidentielle française de 2002. *Revue française de science politique*, 54(1):99, 2004.
[23] Tom Leinster. *Entropy and Diversity: The Axiomatic Approach*. Cambridge University Press, 2021.
[24] Tom Leinster and Christina A. Cobbold. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489, March 2012.
[25] Tom Leinster and Christina A Cobbold. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489, 2012.
[26] Nicholas Mattei and Toby Walsh. A preflib.org retrospective: Lessons learned and new directions. *Trends in Computational Social Choice. AI Access Foundation*, pages 289–309, 2017.
[27] E.C. Pielou. Ecological diversity. *John Wiley & Sons*, 1975.
[28] Ariel D. Procaccia, Jeffrey S. Rosenschein, and Aviv Zohar. On the complexity of achieving proportional representation. *Soc. Choice Welf.*, 30(3):353–362, 2008.
[29] Kunal Relia. Dire committee : Diversity and representation constraints in multiwinner elections. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI'22)*, pages 5143–5149. International Joint Conferences on Artificial Intelligence Organization, 7 2022.
[30] RD Routledge. Diversity indices: which ones are admissible? *Journal of theoretical Biology*, 76(4): 503–515, 1979.
[31] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.
[32] Edward H. Simpson. Measurement of diversity. *Nature*, 163(4148):688–688, 1949.
[33] Andrzej Straszak, Marek Libura, Jarostaw Sikorski, and Dariusz Wagner. Computer-assisted constrained approval voting. *Group Decision and Negotiation*, 2:375–385, 1993.
[34] Clinton Gubong Gassi Takoulo, Mostapha Diss, and Issofa Moyouwou. Combining diversity and excellence in multiwinner elections. Available at SSRN: https://ssrn.com/abstract=4730256 or http://dx.doi.org/10.2139/ssrn.4730256, preprint not peer-reviewed, 2023.
[35] Robert H. Whittaker. Evolution and measurement of species diversity. *Taxon*, 21(2-3):213–251, 1972.

Paula Böhm
TU Clausthal
Institut für Informatik
Clausthal-Zellerfeld, Germany
Email: paula.boehm@tu-clausthal.de

Robert Bredereck
TU Clausthal
Institut für Informatik
Clausthal-Zellerfeld, Germany
Email: robert.bredereck@tu-clausthal.de

Till Fluschnik
Humboldt-Universität zu Berlin
Algorithm Engineering Group
Berlin, Germany
Email: till.fluschnik@hu-berlin.de

Appendix

A Additional Material for Section 3

A.1 Testing the New Index against Adapted Properties

In this section, we adapt properties defined in the literature as desirable for diversity indices and test our new diversity index LC against them. Not all of these properties are necessary in our context, but we look at them to provide some reasons why we call the new index *diversity* index. We consider only the new diversity index here, but we want to stress that not each of the indices that are used in ecology and that we consider satisfies all following adapted properties.

Set Monotonicity. This property defined in [17] states that the diversity should increase if a new species is added to a set of species. We adapt this property by replacing the species with the labels and demanding the following:

A diversity index D satisfies *Set Monotonicity* if, for all elections $\mathcal{E}_1 = (A, C, U, k, L, \lambda)$ and $\mathcal{E}_2 = (A, C, U, k+1, L, \lambda)$ (i.e., only the committee size differs) with $m > k$ and for all committees $S_1 \in \mathcal{R}_{\text{vld}}(\mathcal{E}_1)$ and $S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E}_2)$ with $S_2 = S_1 \cup \{c'\}$ and $n_i(\mathcal{E}_1, S_1) = 0$ with $l_i = \lambda(c')$, it holds that $D(S_1) < D(S_2)$.

LC satisfies this property, as

$$LC(\mathcal{E}_2, S_2) - LC(\mathcal{E}_1, S_1) = \left(\sum_{i=1}^{k} \left((k+2)^{k+2-i} - (k+1)^{k+1-i} \right) \cdot |\sigma_i| \right) + (k+2)^{k+1} > 0.$$

This property is not important in our context, as we search for a committee of a fixed size and, therefore, only need to compare committees with this size based on the same candidate pool.

More Species do not Harm. This property mentioned in [17], [5], and in [15] states that the diversity should not decrease if a new species is added to a set of species with equal frequencies in such a way that the frequencies of all species are equal. We adapt this property by replacing the species with the labels and demanding the following:

A diversity index D satisfies this property if for all elections $\mathcal{E}_1 = (A, C, U, k, L, \lambda)$ with m labels for which $\exists i \in \mathbb{N} : k = m \cdot i$ and for all $\mathcal{E}_2 = (A, C \cup C', U, k+i, L', \lambda')$ with $m+1$ labels, $|C'| = i$ and L', λ' leading to all candidates from C having the same labels as in $\mathcal{E}_1$ and all candidates from C' having the same label l_{m+1} not present in L, it holds that $D(S_1) \leq D(S_2)$ for all $S_1 \in \mathcal{R}_{\text{vld}}(\mathcal{E}_1)$, $S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E}_2)$ with $\forall l \in [m] : n_l(\mathcal{E}_1, S_1) = n_l(\mathcal{E}_2, S_2) = i$ and $n_{m+1}(\mathcal{E}_2, S_2) = i$.

LC satisfies this property, as

$$LC(\mathcal{E}_1, S_1) = \sum_{j=1}^{i} m \cdot (m+1)^{k+1-j} < \sum_{j=1}^{i} (m+1) \cdot (m+2)^{k+i+1-j} = LC(\mathcal{E}_2, S_2)\,.$$

This property is not important in our context, either, for the same reasons as the previous property.

Equal Frequencies are Optimal. A property mentioned in [5] and [15] demands of a diversity index (for a given number of species) that it is maximal if the frequencies of the species are equal. We adapt this property by demanding the following:

A diversity index D satisfies this property if, for all elections $\mathcal{E} = (A, C, U, k, L, \lambda)$ with m labels for which $\exists i \in \mathbb{N} : k = m \cdot i$ or for which $m \geq k$ (let $i = 1$ in the latter case), it holds that $D(S_1)$ is optimal if $S_1 \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ and $\exists L' \subseteq L$: $|L'| = \min\{m, k\}$ and $\forall l_j \in L' : n_j(\mathcal{E}, S_1) = i$.

LC satisfies this property: Let $S_1 \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ be a committee satisfying this condition and $S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ a committee violating this property, i.e., $\exists l \in [m] : n_l(\mathcal{E}, S_2) > i$. S_2 can be transformed into S_1 iteratively by replacing a candidate with a label that occurs more than i times with a candidate with a label that occurs less than i times, which leads to a strictly higher diversity. More formally:

1. Let $l_1 \in [m]$ be a label with $i_1 \coloneqq n_{l_1}(\mathcal{E}, S_2) < i$ and $l_2 \in [m]$ be a label with $i_2 \coloneqq n_{l_2}(\mathcal{E}, S_2) > i$.
2. Let $S_3 \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ be a committee with $n_{l_1}(\mathcal{E}, S_3) = i_1 + 1$, $n_{l_2}(\mathcal{E}, S_3) = i_2 - 1$, and $n_f(\mathcal{E}, S_3) = n_f(\mathcal{E}, S_2)$ for $f \in [m] \setminus \{l_1, l_2\}$.
3. Set $S_2 = S_3$.
4. If there is a $L' \subseteq L$ with $|L'| = m$ so that $\forall l_j \in L' : n_j(\mathcal{E}, S_2) = i$, stop. Otherwise, continue with the first step.

In the second step, the diversity of S_3 is strictly higher than that of S_2 according to LC, with $\eta = \min\{m, k\} + 1$:

$$LC(S_3) - LC(\mathcal{E}, S_2) = \eta^{k-i_1} - \eta^{k+1-i_2} = \eta^{k-i_1}\left(1 - \eta^{1-(i_2-i_1)}\right) > 0 \text{ because } i_2 - i_1 > 1.$$

Clearly, this property is desirable in our context: If it is possible that each label occurs equally often, this should lead to the highest diversity.

Symmetry. This obviously desirable property mentioned in [25] and [15] demands from a diversity index that its value remains the same regardless of the order of the species. When adapting this property by replacing the species with the labels, LC fulfills this property trivially.

Effective Number. A property mentioned in [25] and [30] demands from a diversity index that its value for a candidate set with s species equals s if each species has a frequency of s^{-1}. We adapt his property by demanding the following of a diversity index D: For all elections $\mathcal{E} = (A, C, U, k, L, \lambda)$ with m labels for which $\exists i \in \mathbb{N} : k = m \cdot i$ and $S \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ with $\forall l \in [m] : n_l(\mathcal{E}, S) = i$, it holds that $D(S) = m$.

This property is not fulfilled by LC, as $LC(\mathcal{E}, S) = \sum_{j=1}^{i} m(m+1)^{k+1-j}$.

Because LC does not fulfill this property, it also does not fulfil a different property defined in [30], which demands that the diversity is smaller than the number of species if the species do not have the same frequency. Analogously, the diversity index does not satisfy a property defined in [25] which requires an index's value range to be $\{1, \ldots, s\}$, where s is the number of species (which is m when replacing species with labels).

Are these properties important in our context? It seems that this depends on the situation. It is useful, for example, if it is desirable to derive information about the present species directly from the value of the diversity index (i.e., the effective number of species; for more information see [18], for example). If, on the other hand, the distr vector is given, the number of represented species and which of two given sets is more diverse according to LC is *easy* to see (see Section 4). In such a situation, one could do without this property. In addition, note that the other diversity indices we consider do not satisfy this property, either, Ri being the only exception.

Absent Species. This property mentioned in [25] requires an index's value to remain the same if a new species is added that occurs zero times. We adapt this property by replacing the species with the labels and demanding the following from a diversity index D: For all elections $\mathcal{E}_1 = (A, C, U, k, L, \lambda)$ with m labels and for all $\mathcal{E}_2 = (A, C \cup \{c'\}, U, k, L', \lambda')$ with $c' \notin C$, $m+1$ labels and L', λ' leading to all candidates from C having the same labels as in $\mathcal{E}_1$ and c' having label l_{m+1} not present in L, it holds that $D(S_1) = D(S_2)$ with $S_1 \in \mathcal{R}_{\text{vld}}(\mathcal{E}_1)$,$S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E}_2)$, and $\forall l \in [m] : n_l(\mathcal{E}_1, S_1) = n_l(\mathcal{E}_2, S_2)$ and $n_{m+1}(\mathcal{E}_2, S_2) = 0$.

LC does not fulfill this property: Consider, for example, as $\mathcal{E}_1$ an election with $m = 2$ and each label occurring twice (i.e., $k = 4$), then it holds that $LC(\mathcal{E}_2, S_2) - LC(\mathcal{E}_1, S_1) = 424 > 0$. This property is not important in our context for the same reasons as for *Set Monotonicity*.

B Additional Material for Section 4

B.1 Proof of Observation 1

We observed this programmatically: For each committee size $k \in \{1, \ldots, 7\}$, we iterated over each possible number m of labels that can occur in the committee, i.e., $m \in [k]$, as each candidate can introduce at most one label. For each such combination of k and m, we looked at all possible distr vectors and computed and compared the diversity indices of interest, exploiting the fact that the naming and ordering of the candidates and agents as well as their votes do not matter for the diversity indices at hand. □

B.2 Proof of Observation 2

Diversity indices Sh and Si: Consider an election $\mathcal{E}$ with $k = 10$ and $m = 3$ labels, and two committees $S_1, S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ with $p_1(\mathcal{E}, S_1) = p_2(\mathcal{E}, S_1) = 0.1, p_3(\mathcal{E}, S_1) = 0.8$, and $p_1(\mathcal{E}, S_2) = 0, p_2(\mathcal{E}, S_2) = p_3(\mathcal{E}, S_2) = 0.5$. It holds that $\text{distr}(\mathcal{E}, S_1)_1 = 0 < 1 = \text{distr}(S_2)_1$, but $Sh(S_1) \approx 0.64 < Sh(S_2) \approx 0.69$, and $Si(S_1) = -0.66 < Si(S_2) = -0.5$.

Diversity index Ri: This follows directly from the definition of Ri as $m - \text{distr}(\mathcal{E}, S)_1$.

Diversity index LC: As $\text{distr}(\mathcal{E}, S_1) \neq \text{distr}(\mathcal{E}, S_2)$, it follows from Observation 6 that $LC(\mathcal{E}, S_1) \neq LC(\mathcal{E}, S_2)$. Thus, it follows from Observation 5 (with $\rho_1 = 1$) that $LC(S_1) > LC(S_2)$. □

B.3 Proof of Observation 3

In the following, let $n'_l = n_l(\mathcal{E}, S')$, $n''_l = n_l(\mathcal{E}, S'')$ for $l \in [m]$, $\text{distr}' = \text{distr}(\mathcal{E}, S')$, $\text{distr}'' = \text{distr}(\mathcal{E}, S'')$. For the meaning of i, j, S' and S'' see the definition of *Occurrence Balancing*.

Diversity index Ri: If $n'_i = 0$ and thus $\text{distr}'_1 > \text{distr}''_1$, it holds that $Ri(\mathcal{E}, S') = m - \text{distr}'_1 < m - \text{distr}''_1 = Ri(\mathcal{E}, S'')$. Otherwise, i.e., if $\text{distr}'_1 = \text{distr}''_1$, $Ri(\mathcal{E}, S') = Ri(\mathcal{E}, S'') = m - \text{distr}'_1$. Hence, $Ri(S') \leq Ri(S'')$.

Diversity index Sh: Let $M'_R = \{i \in [m] : p_i(\mathcal{E}, S') > 0\}$ and $M''_R = \{i \in [m] : p_i(\mathcal{E}, S'') > 0\}$. It holds that $Sh(\mathcal{E}, S'') - Sh(\mathcal{E}, S') > 0 \Leftrightarrow \sum_{p \in M'_R} \frac{n'_p}{k} \cdot \log\left(\frac{n'_p}{k}\right) - \sum_{p \in M''_R} \frac{n''_p}{k} \cdot \log\left(\frac{n''_p}{k}\right) > 0 \Leftrightarrow \sum_{p \in M_R} n'_p \cdot \log\left(n'_p\right) - \sum_{p \in M_R} n''_p \cdot \log\left(n''_p\right) > 0$.
If $n'_i = 0$, this is equivalent to $n'_j \cdot \log\left(n'_j\right) - \left(n'_j - 1\right) \cdot \log\left(n'_j - 1\right) > 0$, which is true because log is strictly monotonically increasing. Otherwise, i.e., if $n'_i > 0$, it is equivalent to $n'_j \cdot \log\left(n'_j\right) - \left(n'_j - 1\right) \cdot \log\left(n'_j - 1\right) + n'_i \cdot \log(n'_i) - (n'_i + 1) \cdot \log(n'_i + 1) > 0$, which we prove in the following:

Let $f(x) = x\log(x)$ with $f'(x) = 1 + \log(x)$ which is strictly monotonically increasing. According to the mean value theorem, $\exists c_1 \in \left(n'_j - 1, n'_j\right), c_2 \in (n'_i, n'_i + 1)$ such that $f\left(n'_j\right) - f\left(n'_j - 1\right) = f'(c_1) = 1 + \log(c_1)$ and $f(n'_i + 1) - f(n'_i) = f'(c_2) = 1 + \log(c_2)$. Thus, the former inequality is equivalent to $1 + \log(c_1) - 1 - \log(c_2) = \log(c_1) - \log(c_2) > 0$, which is true due to $c_1 > c_2$.

Diversity index Si**:** It holds that $Si(\mathcal{E}, S'') - Si(\mathcal{E}, S') > 0 \Leftrightarrow \sum_{p\in[m]} \left(n''_p\right)^2 - \left(n'_p\right)^2 < 0 \Leftrightarrow \left(n'_j - 1\right)^2 - \left(n'_j\right)^2 + \left(n'_i + 1\right)^2 - \left(n'_i\right)^2 < 0 \Leftrightarrow 2 + 2\left(n'_i - n'_j\right) < 0$. The latter is true, as $n'_i + 1 < n'_j$ and thus $n'_i - n'_j \leq -2$.

Diversity index LC**:** Let $\delta \coloneqq n'_j - n'_i$ and $\eta = \min\{m, k\} + 1$. It holds that

$$LC\left(S''\right) - LC\left(S'\right) > 0 \Leftrightarrow \eta^{k+1-n'_i-1} - \eta^{k+1-n'_j} = \eta^{k-n'_i} - \eta^{k+1-n'_i-\delta} \quad > 0 \Leftrightarrow 1 - \eta^{1-\delta} > 0$$

which is true, as $1 - \delta < 0$. □

B.4 Proof of Observation 4

In the following, let $\alpha \coloneqq n_i(\mathcal{E}, S)$ and $d_2 \coloneqq \lfloor \frac{d}{2} \rfloor$.

Diversity index Ri**:** If $n_i(\mathcal{E}, S) \geq 1$, it holds that $\text{distr}(\mathcal{E}, S)_1 = \text{distr}(\mathcal{E}, S_{(i,j)})_1 = \text{distr}(\mathcal{E}, S_{(k,l)})_1$ and therefore $Ri\left(S_{(i,j)}\right) = m - \text{distr}(\mathcal{E}, S)_1 = Ri\left(S_{(k,l)}\right)$.

Diversity index Si**:** It holds that

$$\begin{aligned}
& Si\left(S_{(i,j)}\right) - Si(S) \\
= & \left(-\left(\alpha + d_2\right)^2 - \left(\alpha + d - d_2\right)^2 + \alpha^2 + \left(\alpha + d\right)^2\right)/k^2 \\
= & \left(-\alpha^2 - 2\alpha d_2 - d_2^2 - \alpha^2 - 2\alpha\left(d - d_2\right) - \left(d - d_2\right)^2 + \alpha^2 + \alpha^2 + 2\alpha d + d^2\right)/k^2 \\
= & \left(-d_2^2 - \left(d - d_2\right)^2 + d^2\right)/k^2
\end{aligned}$$

and analogously $Si\left(S_{(k,l)}\right) - Si(S) = \left(-d_2^2 - \left(d - d_2\right)^2 + d^2\right)/k^2$. Hence, $Si\left(S_{(i,j)}\right) - Si\left(S_{(k,l)}\right) = Si\left(S_{(i,j)}\right) - Si(S) - \left(Si\left(S_{(k,l)}\right) - Si(S)\right) = 0$.

Diversity index LC**:** This follows directly from LC being obvious with $f(l) = 1$ (see Observation 5) because $\text{rdistr}\left(\mathcal{E}, S_{(i,j)}, S_{(k,l)}\right)_1 = \text{distr}\left(\mathcal{E}, S_{(i,j)}\right)_{\alpha+1} < \text{rdistr}\left(\mathcal{E}, S_{(k,l)}, S_{(i,j)}\right)_1 = \text{distr}\left(\mathcal{E}, S_{(k,l)}\right)_{\alpha+1}$.

Diversity index Sh**:** Consider the function

$$f(x) = \frac{-\left(x + d_2\right)\log\left(x + d_2\right) - \left(x + d - d_2\right)\log\left(x + d - d_2\right) + x\log(x) + \left(x + d\right)\log\left(x + d\right)}{k}.$$

Hence, $f(\alpha) = Sh\left(S_{(i,j)}\right) - Sh(S)$ and $f(n_k(\mathcal{E}, S)) = Sh\left(S_{(k,l)}\right) - Sh(S)$. The derivative is

$$f'(x) = (\log(x) + \log(d + x) - \log(d - d_2 + x) - \log(d_2 + x))/k = \log\left(\frac{x\left(d + x\right)}{\left(d - d_2 + x\right)\left(d_2 + x\right)}\right)/k.$$

As

$$\frac{x\left(d + x\right)}{\left(d - d_2 + x\right)\left(d_2 + x\right)} = \frac{xd + x^2}{dd_2 + dx - d_2^2 - d_2x + d_2x + x^2} = \frac{xd + x^2}{xd + x^2 + dd_2 - d_2^2} < 1$$

due to $d > d_2$, it holds that $f'(x) < 0$ for $x \geq 0$. Hence, $f(x)$ is strictly monotonically decreasing for $x \geq 0$ and $f(\alpha) = Sh\left(S_{(i,j)}\right) - Sh(S) > f(n_k(\mathcal{E}, S)) = Sh\left(S_{(k,l)}\right) - Sh(S) \Leftrightarrow Sh\left(S_{(i,j)}\right) > Sh(S_{(k,l)})$. □

B.5 Proof of Observation 5

In the following, let $\text{rdistr}^{(1)} = \text{rdistr}(\mathcal{E}, S_1, S_2)$, $\text{rdistr}^{(2)} = \text{rdistr}(\mathcal{E}, S_2, S_1)$, $\text{distr}^{(1)} = \text{distr}(\mathcal{E}, S_1)$, $\text{distr}^{(2)} = \text{distr}(\mathcal{E}, S_2)$, ρ the vector of the elements of $I_R(\mathcal{E}, S_1, S_2)$ in ascending order, and $\eta := \min\{m, k\} + 1$.

Diversity indices Sh and Si: These two indices violate this property. A counterexample using an election $\mathcal{E}$ with $k = 8$ and $m = 6$ is the following: Consider $\text{rdistr}^{(1)} = (2, 0, 4, 0)$ and $\text{rdistr}^{(2)} = (0, 5, 0, 1)$ with $\rho_1 = 1$, $\rho_2 = 2$, $\rho_3 = 3$, $\rho_4 = 4$. Therefore, $Si(\mathcal{E}, S_1) = -\frac{16}{64} < Si(\mathcal{E}, S_2) = -\frac{14}{64}$ and $Sh(\mathcal{E}, S_1) \approx 1.39 < Sh(\mathcal{E}, S_2) \approx 1.67$. Thus, both indices categorize S_2 as more diverse. This makes it a counterexample for $l' = 2$ and $l' = 4$.

For $l' = 1$, $l' = 3$, and Si, the following is a counterexample: Consider $\text{rdistr}^{(1)} = (1, 4, 0, 1)$ and $\text{rdistr}^{(2)} = (2, 0, 4, 0)$ with $\rho_1 = 1, \rho_2 = 2, \rho_3 = 3, \rho_4 = 5$. It holds that $Si(\mathcal{E}, S_1) = -\frac{20}{64} < Si(\mathcal{E}, S_2) = -\frac{16}{64}$.

For $l' = 1$, $l' = 3$, and Sh, the following is a counterexample: Consider $\text{rdistr}^{(1)} = (0, 5, 0, 1)$ and $\text{rdistr}^{(2)} = (1, 2, 3, 0)$ with $\rho_1 = 1, \rho_2 = 2, \rho_3 = 3, \rho_4 = 4$. It holds that $Sh(\mathcal{E}, S_1) \approx -1.67 < Sh(\mathcal{E}, S_2) = -1.56$.

Diversity index Ri: Ri is 1-obvious[7]. It follows from the definition of Ri that $Ri(\mathcal{E}, S_1) = \sum_{i=1}^{k} \text{distr}_{i+1}^{(1)} = m - \text{distr}_1^{(1)}$ and analogously for $Ri(\mathcal{E}, S_2)$. As we assume that the corresponding candidate sets are not equally diverse with respect to Ri, $\text{distr}_1^{(1)} \neq \text{distr}_1^{(2)}$ and $\rho_1 = 1$ holds. Hence, $\text{rdistr}_1^{(1)} < \text{rdistr}_1^{(2)} \Leftrightarrow \text{distr}_1^{(1)} < \text{distr}_1^{(2)} \Leftrightarrow m - \text{distr}_1^{(1)} > m - \text{distr}_1^{(2)} \Leftrightarrow Ri(\mathcal{E}, S_1) > Ri(\mathcal{E}, S_2)$.

Diversity index LC: LC is 1-obvious: We first show that $LC(\mathcal{E}, S_1) > LC(\mathcal{E}, S_2) \Rightarrow \text{rdistr}_1^{(1)} < \text{rdistr}_1^{(2)}$. For this, suppose that $LC(\mathcal{E}, S_1) > LC(\mathcal{E}, S_2)$ and $\text{rdistr}_1^{(1)} > \text{rdistr}_1^{(2)}$, and let $r = \rho_1$ and $\eta = \min\{m, k\} + 1$.

Then it holds that $\forall i \in [r-1] : \text{distr}_i^{(1)} = \text{distr}_i^{(2)} \wedge \sigma(\mathcal{E}, S_1)_i = \sigma(\mathcal{E}, S_2)_i$ and thus $\sigma(\mathcal{E}, S_1)_r < \sigma(\mathcal{E}, S_2)_r$ because $\text{rdistr}_1^{(1)} > \text{rdistr}_1^{(2)} \Leftrightarrow \text{distr}_r^{(1)} > \text{distr}_r^{(2)}$, which means that S_1 has more labels occurring $r-1$ times than S_2 and thus fewer labels occurring at least r times than S_2. It holds that

$$LC(\mathcal{E}, S_1) > LC(\mathcal{E}, S_2) \Leftrightarrow \sum_{i=1}^{k} \eta^{k+1-i} \cdot (|\sigma(\mathcal{E}, S_1)_i| - |\sigma(\mathcal{E}, S_2)_i|) > 0$$

$$\Leftrightarrow \eta^{k+1-r} \cdot (|\sigma(\mathcal{E}, S_1)_r| - |\sigma(\mathcal{E}, S_2)_r|) + \sum_{i=r+1}^{k} \eta^{k+1-i} \cdot (|\sigma(\mathcal{E}, S_1)_i| - |\sigma(\mathcal{E}, S_2)_i|) > 0. \quad (1)$$

As $\sigma(\mathcal{E}, S_1)_r < \sigma(\mathcal{E}, S_2)_r$ and $\forall i \in [k] : \eta \geq \sigma(\mathcal{E}, S')_i \geq 0$ for $S' \in \{S_1, S_2\}$, it follows (with the help of the geometric series formula) that

$$(\eta+1)^{k+1-r} \cdot (|\sigma(\mathcal{E}, S_1)_r| - |\sigma(\mathcal{E}, S_2)_r|) + \sum_{i=r+1}^{k} (\eta+1)^{k+1-i} \cdot (|\sigma(\mathcal{E}, S_1)_i| - |\sigma(\mathcal{E}, S_2)_i|)$$

$$\leq -(\eta+1)^{k+1-r} + \sum_{i=r+1}^{k} \eta \cdot (\eta+1)^{k+1-i} < -(\eta+1)^{k+1-r} + \eta + 1 + \sum_{i=r+1}^{k} \eta \cdot (\eta+1)^{k+1-i}$$

$$= -(\eta+1)^{k+1-r} + \eta + 1 - (\eta+1) + (\eta+1)^{k+1-r} = 0 \quad \lightning \text{ to (1)}$$

For $\text{rdistr}_1^{(1)} < \text{rdistr}_1^{(2)} \Rightarrow LC(\mathcal{E}, S_1) > LC(\mathcal{E}, S_2)$, see the proof for LC for Observation 6. □

[7] We write 1-obvious and l-obvious short for obvious with $f(l) = 1$ for all $l \in \mathbb{N}$ and $f(l) = l$ for all $l \in \mathbb{N}$, respectively.

B.6 Proof of Observation 6

Diversity index Sh**:** Consider an election with $m = 5$ labels and $k = 8$, and two committees $S_1, S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ with $\text{distr}(\mathcal{E}, S_1) = (1, 0, 4, 0, 0)$ and $\text{distr}(\mathcal{E}, S_2) = (0, 4, 0, 0, 1)$. It holds that $\text{distr}(\mathcal{E}, S_1) \neq \text{distr}(\mathcal{E}, S_2)$, but

$$Sh(\mathcal{E}, S_1) = -4 \cdot \frac{2}{8} \cdot \log\left(\frac{2}{8}\right) = -\left(\log(2) - \log(8)\right) = -\left(\frac{1}{2}\left(\log(2) + \log(2)\right) - \log(8)\right)$$
$$= -\left(\frac{1}{2}\log(4) - \log(8)\right) = -\left(4 \cdot \frac{1}{8}\log\left(\frac{1}{8}\right) + \frac{1}{2}\log\left(\frac{4}{8}\right)\right) = Sh(\mathcal{E}, S_2)\,.$$

Diversity index Si**:** Consider an election with $m = 6$ labels and $k = 8$, and two committees $S_1, S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ with $\text{distr}(\mathcal{E}, S_1) = (1, 2, 3, 0)$ and $\text{distr}(\mathcal{E}, S_2) = (0, 5, 0, 1)$. It holds that $Si(\mathcal{E}, S_1) = -\frac{1}{64}(2 + 3 \cdot 4) = -\frac{1}{64}(5 + 9) = Si(\mathcal{E}, S_2)$, although $\text{distr}(\mathcal{E}, S_1) \neq \text{distr}(\mathcal{E}, S_2)$.

Diversity index Ri**:** Consider an election with $m = 2$ labels and $k = 4$, and two committees $S_1, S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ with $\text{distr}(\mathcal{E}, S_1) = (0, 1, 0, 1)$ and $\text{distr}(\mathcal{E}, S_2) = (0, 0, 2, 0)$. It holds that $Ri(\mathcal{E}, S_1) = 2 = Ri(\mathcal{E}, S_2)$, although $\text{distr}(\mathcal{E}, S_1) \neq \text{distr}(\mathcal{E}, S_2)$.

Diversity index LC**:** Clearly, if $\text{distr}(\mathcal{E}, S_1) = \text{distr}(\mathcal{E}, S_2)$, it holds that $LC(\mathcal{E}, S_1) = LC(\mathcal{E}, S_2)$. Next, we assume that $\text{distr}(\mathcal{E}, S_1) \neq \text{distr}(\mathcal{E}, S_2)$. Thus, $\text{rdistr}^{(1)} = \text{rdistr}(\mathcal{E}, S_1, S_2)$ and $\text{rdistr}^{(2)} = \text{rdistr}(\mathcal{E}, S_2, S_1)$ have a length $l \geq 1$. Assume, w.l.o.g., $\text{rdistr}_1^{(1)} < \text{rdistr}_1^{(2)}$ and let $r = \rho_1$ and $\eta = \min\{m, k\} + 1$. Thus, $\forall i \in \{1, \ldots, r-1\} : \text{distr}_i^{(1)} = \text{distr}_i^{(2)} \wedge \sigma(\mathcal{E}, S_1)_i = \sigma(\mathcal{E}, S_2)_i$ and hence $\sigma(\mathcal{E}, S_1)_r > \sigma(\mathcal{E}, S_2)_r$ because $\text{rdistr}_1^{(1)} < \text{rdistr}_1^{(2)} \Leftrightarrow \text{distr}_r^{(1)} < \text{distr}_r^{(2)}$, which means that S_2 has more labels occurring $r-1$ times than S_1 and thus fewer labels occurring at least r times than S_1. Based on this, it holds that

$$LC(\mathcal{E}, S_1) - LC(\mathcal{E}, S_2) = \sum_{i=1}^{k} \eta^{k+1-i} \cdot \left(|\sigma(\mathcal{E}, S_1)_i| - |\mathcal{E}, \sigma(S_2)_i|\right)$$
$$= \eta^{k+1-r} \cdot \left(|\sigma(\mathcal{E}, S_1)_r| - |\sigma(\mathcal{E}, S_2)_r|\right) + \sum_{i=r+1}^{k} \eta^{k+1-i} \cdot \left(|\sigma(\mathcal{E}, S_1)_i| - |\sigma(\mathcal{E}, S_2)_i|\right).$$

As $\sigma(\mathcal{E}, S_1)_r > \sigma(\mathcal{E}, S_2)_r$ and $\forall i \in \{1, \ldots, k\} : \eta - 1 \geq \sigma(\mathcal{E}, S')_r \geq 0$ for $S' \in \{S_1, S_2\}$, it follows (with the help of the geometric series formula) that

$$LC(\mathcal{E}, S_1) - LC(\mathcal{E}, S_2) \geq \eta^{k+1-r} - \sum_{i=r+1}^{k} (\eta - 1) \cdot \eta^{k+1-i}$$
$$> \eta^{k+1-r} - \eta - \sum_{i=r+1}^{k} (\eta - 1) \cdot \eta^{k+1-i} = \eta^{k+1-r} - \eta + \eta - \eta^{k+1-r} = 0.$$

Thus, $LC(\mathcal{E}, S_2) < LC(\mathcal{E}, S_1)$. □

B.7 Proof of Theorem 1

In the following, let $\text{rdistr}^{(1)} = \text{rdistr}(\mathcal{E}, S_1, S_2)$, $\text{rdistr}^{(2)} = \text{rdistr}(\mathcal{E}, S_2, S_1)$. In addition, let $\mathcal{E}$ be an arbitrary, but fixed, election, and $S_1, S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E})$.

Let D_d be a diversity index satisfying *Distribution Equivalence*, *Obviousness*, and *Present Label Maximization*. If $LC(\mathcal{E}, S_1) = LC(\mathcal{E}, S_2)$ ($D_d(\mathcal{E}, S_1) = D_d(\mathcal{E}, S_2)$), it holds that $D_d(\mathcal{E}, S_1) = D_d(\mathcal{E}, S_2)$ ($LC(\mathcal{E}, S_1) = LC(\mathcal{E}, S_2)$), as this holds for both diversity indices if and only if $\text{distr}(\mathcal{E}, S_1) = \text{distr}(\mathcal{E}, S_2)$ because of *Distribution Equivalence*.

If, however, $\text{distr}(\mathcal{E}, S_1) \neq \text{distr}(\mathcal{E}, S_2)$ we show that D_d satisfies *Obviousness* with $f(l) = 1$ if the length l of the rdistr vectors is at least four and, otherwise, with such $f(l) = l'$ which are also *valid* for LC. Thus, we show that $LC(\mathcal{E}, S_1) > LC(\mathcal{E}, S_2)$ if and only if $D_d(\mathcal{E}, S_1) > D_d(\mathcal{E}, S_2)$. For this, we will use that the following holds with l being the length of the rdistr vectors of S_1 and S_2:

$$\sum_{i=1}^{l} \text{rdistr}_i^{(1)} = \sum_{i=1}^{l} \text{rdistr}_i^{(2)} \tag{2}$$

$$\sum_{i=1}^{l} \rho_i \cdot \text{rdistr}_i^{(1)} = \sum_{i=1}^{l} \rho_i \cdot \text{rdistr}_i^{(2)} \tag{3}$$

First, note that $l < 3$ is not possible: Assume that $l = 1$ is possible, then there is exactly one index r_1 for which $\text{distr}(\mathcal{E}, S_1)_{r_1} \neq \text{distr}(\mathcal{E}, S_2)_{r_1}$ and let w.l.o.g. $\text{distr}(\mathcal{E}, S_1)_{r_1} > \text{distr}(\mathcal{E}, S_2)_{r_1}$. However, there must be an r_2 with $\text{distr}(\mathcal{E}, S_1)_{r_2} < \text{distr}(\mathcal{E}, S_2)_{r_2}$ because of Eq. (2), i.e., l needs to be at least two. W.l.o.g., let $r_1 < r_2$. Assume, that $l = 2$ is possible and thus $d := \text{distr}(\mathcal{E}, S_1)_{r_1} - \text{distr}(\mathcal{E}, S_2)_{r_1} = \text{distr}(\mathcal{E}, S_2)_{r_2} - \text{distr}(\mathcal{E}, S_1)_{r_2}$. Then, it holds that $\sum_{i \in [k+1]} (i-1) \cdot (\text{distr}(\mathcal{E}, S_1)_i - \text{distr}(\mathcal{E}, S_2)_i) = (r_1 - 1) \cdot d - (r_2 - 1) \cdot d = d \cdot (r_1 - r_2) \neq 0$. This is a contradiction to Eq. (3). Thus, l is at least three.

Next, we assume that $l \geq 4$ and show that D_d satisfies *Obviousness* only with $f(l) = 1$ (like LC does). As a counterexample for $f(l) = l' \in \{2, \ldots, l-1\}$ consider an election $\mathcal{E}_1$ with $m = l - 2$ labels and $k = \sum_{j=2}^{l-1} j - 1$ and consider $S_1 \in \mathcal{R}_{\text{vld}}(\mathcal{E}_1)$ with

$$\begin{aligned} \text{rdistr}(\mathcal{E}_1, S_1, S_2)_1 &= 0 \\ \forall j \in \{2, \ldots, l-1\} : \text{rdistr}(\mathcal{E}_1, S_1, S_2)_j &= 1 \\ \text{rdistr}(\mathcal{E}_1, S_1, S_2)_l &= 0 \end{aligned}$$

and $S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E}_1)$ with

$$\begin{aligned} \text{rdistr}(\mathcal{E}_1, S_2, S_1)_1 &= l - 2 - 1 \\ \forall j \in \{2, \ldots, l-1\} : \text{rdistr}(\mathcal{E}_1, S_2, S_1)_j &= 0 \\ \text{rdistr}(\mathcal{E}_1, S_2, S_1)_l &= 1 \end{aligned}$$

with $\forall j \in [l-1] : \rho_j = j$, and $\rho_l = k$. Thus, $D_d(\mathcal{E}_1, S_1) > D_d(\mathcal{E}_1, S_2)$ and $LC(\mathcal{E}_1, S_1) > LC(\mathcal{E}_1, S_2)$ (because both indices satisfy *Present Label Maximization*) and $l' \notin \{2, \ldots, l-1\}$.
For a counterexample for $f(l) = l' = l$ consider an election $\mathcal{E}_2$ with $m = l$ labels and $k = \left(\sum_{j=3}^{l-2} j - 1\right) + 3\,(l-2)$ and consider $S_1 \in \mathcal{R}_{\text{vld}}(\mathcal{E}_2)$ with

$$\begin{aligned} \text{rdistr}(\mathcal{E}_2, S_1, S_2)_1 &= 0 \\ \text{rdistr}(\mathcal{E}_2, S_1, S_2)_2 &= l - 1 \\ \forall j \in \{3, \ldots, l-2\} : \text{rdistr}(\mathcal{E}_2, S_1, S_2)_j &= 0 \\ \text{rdistr}(\mathcal{E}_2, S_1, S_2)_{l-1} &= 0 \\ \text{rdistr}(\mathcal{E}_2, S_1, S_2)_l &= 1 \end{aligned}$$

and $S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E}_2)$ with

$$\begin{aligned} \text{rdistr}(\mathcal{E}_2, S_2, S_1)_1 &= 1 \\ \text{rdistr}(\mathcal{E}_2, S_2, S_1)_2 &= 0 \\ \forall j \in \{3, \ldots, l-2\} : \text{rdistr}(\mathcal{E}_2, S_2, S_1)_j &= 1 \\ \text{rdistr}(\mathcal{E}_2, S_2, S_1)_{l-1} &= 3 \\ \text{rdistr}(\mathcal{E}_2, S_2, S_1)_l &= 0 \end{aligned}$$

with $\rho_j = j \; \forall j \in [l-1]$, and $\rho_l = k - (l-1)$. Thus, $D_d(\mathcal{E}_2, S_1) > D_d(\mathcal{E}_2, S_2)$ and $LC(\mathcal{E}_2, S_1) > LC(\mathcal{E}_2, S_2)$ (because both indices satisfy *Present Label Maximization*) and $l' \notin \{2, l\}$. Therefore, D_d satisfies *Obviousness* only with $f(l) = 1$ when $l \geq 4$.

Next, we consider $l = 3$. Assume that $f(l) = l' \neq 1$ and w.l.o.g. $D_d(\mathcal{E}, S_1) > D_d(\mathcal{E}, S_2)$. As D_d satisfies *Present Label Maximization*, it needs to hold that $\text{rdistr}_1^{(1)} < \text{rdistr}_1^{(2)} \Rightarrow \text{rdistr}_{l'}^{(1)} < \text{rdistr}_{l'}^{(2)}$. We show that $\text{rdistr}_1^{(1)} < \text{rdistr}_1^{(2)} \Rightarrow \text{rdistr}_3^{(1)} < \text{rdistr}_3^{(2)}$ and $\text{rdistr}_1^{(1)} < \text{rdistr}_1^{(2)} \Rightarrow \text{rdistr}_2^{(1)} > \text{rdistr}_2^{(2)}$.

For this, assume that $\text{rdistr}_1^{(1)} < \text{rdistr}_1^{(2)}$ and let $\rho_2 = \rho_1 + \delta_1$ and $\rho_3 = \rho_1 + \delta_1 + \delta_2$ with $\delta_1, \delta_2 > 0$. Because of Eq. (3), it needs to hold that

$$\begin{aligned}
&\rho_1 \left(\text{rdistr}_1^{(1)} + \text{rdistr}_2^{(1)} + \text{rdistr}_3^{(1)}\right) + \delta_1 \left(\text{rdistr}_2^{(1)} + \text{rdistr}_3^{(1)}\right) + \delta_2 \cdot \text{rdistr}_3^{(1)} \\
=&\rho_1 \left(\text{rdistr}_1^{(2)} + \text{rdistr}_2^{(2)} + \text{rdistr}_3^{(2)}\right) + \delta_1 \left(\text{rdistr}_2^{(2)} + \text{rdistr}_3^{(2)}\right) + \delta_2 \cdot \text{rdistr}_3^{(2)}
\end{aligned}$$

and, therefore, it needs to hold because of Eq. (2) that

$$\delta_1 \left(\text{rdistr}_2^{(1)} + \text{rdistr}_3^{(1)}\right) + \delta_2 \cdot \text{rdistr}_3^{(1)} = \delta_1 \left(\text{rdistr}_2^{(2)} + \text{rdistr}_3^{(2)}\right) + \delta_2 \cdot \text{rdistr}_3^{(2)} \tag{4}$$

Assume that $\text{rdistr}_3^{(1)} > \text{rdistr}_3^{(2)}$. Because of Eq. (2) and $\text{rdistr}_1^{(1)} < \text{rdistr}_1^{(2)}$, it holds that $\text{rdistr}_2^{(2)} + \text{rdistr}_3^{(2)} < \text{rdistr}_2^{(1)} + \text{rdistr}_3^{(1)}$ and thus

$$\delta_1 \left(\text{rdistr}_2^{(2)} + \text{rdistr}_3^{(2)}\right) + \delta_2 \cdot \text{rdistr}_3^{(2)} < \delta_1 \left(\text{rdistr}_2^{(1)} + \text{rdistr}_3^{(1)}\right) + \delta_2 \cdot \text{rdistr}_3^{(1)}$$

which is a contradiction to Eq. (4). Thus, $\text{rdistr}_1^{(1)} < \text{rdistr}_1^{(2)} \Rightarrow \text{rdistr}_3^{(1)} < \text{rdistr}_3^{(2)}$ and, based on this and Eq. (2), $\text{rdistr}_1^{(1)} < \text{rdistr}_1^{(2)} \Rightarrow \text{rdistr}_2^{(1)} > \text{rdistr}_2^{(2)}$. Therefore, D_d satisfies *Obviousness* with $l' = 1$ or $l' = 3$.

Finally, we show that this also holds for LC by showing that $\text{rdistr}_3^{(1)} < \text{rdistr}_3^{(2)} \Rightarrow \text{rdistr}_1^{(1)} < \text{rdistr}_1^{(2)}$. Assume that $\text{rdistr}_3^{(1)} < \text{rdistr}_3^{(2)}$, but $\text{rdistr}_1^{(1)} > \text{rdistr}_1^{(2)}$. Because of Eq. (4), it follows that $\text{rdistr}_2^{(1)} > \text{rdistr}_2^{(2)}$. Let $\delta_1' := \text{rdistr}_1^{(1)} - \text{rdistr}_1^{(2)}$, $\delta_2' := \text{rdistr}_2^{(1)} - \text{rdistr}_2^{(2)}$. It holds that $\delta_1', \delta_2' > 0$ and, because of Eq. (2), $\text{rdistr}_3^{(2)} = \text{rdistr}_3^{(1)} + \delta_1' + \delta_2'$. Thus, it follows from Eq. (4) that it needs to hold that

$$\begin{aligned}
&\delta_1 \left(\text{rdistr}_2^{(1)} + \text{rdistr}_3^{(1)} - \text{rdistr}_2^{(2)} - \text{rdistr}_3^{(2)}\right) + \delta_2 \cdot \left(\text{rdistr}_3^{(1)} - \text{rdistr}_3^{(2)}\right) \\
&= \delta_1 \left(\text{rdistr}_2^{(1)} + \text{rdistr}_3^{(1)} - \text{rdistr}_2^{(2)} - \text{rdistr}_3^{(1)} - \delta_1' - \delta_2'\right) + \delta_2 \cdot \left(\text{rdistr}_3^{(1)} - \text{rdistr}_3^{(1)} - \delta_1' - \delta_2'\right) \\
&= \delta_1 \left(-\delta_1'\right) + \delta_2 \left(-\delta_1' - \delta_2'\right) = 0,
\end{aligned}$$

a contradiction, as $\delta_1' > 0, \delta_1' + \delta_2' > 0, \delta_1 > 0$, and $\delta_2 > 0$. □

B.8 Further Information

Next, we show the connection between LC and Ri: If $\text{distr}(\mathcal{E}, S_1) = \text{distr}(\mathcal{E}, S_2)$, $LC(\mathcal{E}, S_2) = LC(\mathcal{E}, S_1)$ (because LC satisfies *Distribution Equivalence*) and $Ri(\mathcal{E}, S_1) = Ri(\mathcal{E}, S_2)$ clearly holds. On the other hand, if $LC(\mathcal{E}, S_1) > LC(\mathcal{E}, S_2)$, Ri either classifies both as equally diverse or S_1 as more diverse, because it is obvious with $f(l) = 1$, as is LC. Thus, we have the following:

Corollary 3. *For each election $\mathcal{E}$ and $S_1, S_2 \in \mathcal{R}_{\text{vld}}(\mathcal{E})$, it holds that $LC(\mathcal{E}, S_1) \geq LC(\mathcal{E}, S_2) \Rightarrow Ri(\mathcal{E}, S_1) \geq Ri(\mathcal{E}, S_2)$.*

C Additional Material for Section 5.1

C.1 Proof of Observation 7

The following algorithm yields a most diverse committee:

1. Start with an empty candidate set $C' = \emptyset$.
2. Pick an $l' \in \arg\min_{l \in L'} n_l(\mathcal{E}, C')$ with $L' = \{l \in [m] : n_l(\mathcal{E}, C \setminus C') > 0\}$ (i.e. a label is in L' if it is assigned to at least one candidate not chosen yet).
3. Add a $c \in C_{label}(\mathcal{E}, C \setminus C', l')$ to C'.
4. If $|C'| = k$, stop. Otherwise, go to step 2.

Clearly, this returns a committee C' satisfying the following property: $\forall\, (l, l') \in [m] \times [m] : n_{l'}(\mathcal{E}, C') + 1 < n_l(\mathcal{E}, C') \Rightarrow n_{l'}(\mathcal{E}, C \setminus C') = 0$ (otherwise it would contradict step 2 and 3).

In the following, let $\mathrm{distr}' = \mathrm{distr}(\mathcal{E}, C')$. First, we show that all committees from $\mathcal{R}_{\mathrm{vld}}(\mathcal{E})$ fulfilling the above property have the same distr vector and therefore the same diversity according to each of the indices at hand. Assume this is not the case, i.e., there exists a $C'' \in \mathcal{R}_{\mathrm{vld}}(\mathcal{E})$ with $\mathrm{distr}'' \coloneqq \mathrm{distr}(\mathcal{E}, C'') \neq \mathrm{distr}'$ that fulfills the property. Thus, there are $l \in [m]$, $l' \in [m] \setminus \{l\}$ and $d, d' \in \mathbb{N}$ such that $n_l(\mathcal{E}, C') = n_l(\mathcal{E}, C'') + d$ and $n_{l'}(\mathcal{E}, C') + d' = n_{l'}(\mathcal{E}, C'')$ and, hence, $n_l(\mathcal{E}, C \setminus C'') > 0$. We make a case distinction as to how much the number of occurrences of l and l' differ in C'—note that $n_l(\mathcal{E}, C') > n_{l'}(\mathcal{E}, C') + 1$ is not possible because C' satisfies the property—, each of which leads to a contradiction:

- $n_l(\mathcal{E}, C') \leq n_{l'}(\mathcal{E}, C')$: As the number of l is smaller in C'' than in C', $n_l(\mathcal{E}, C \setminus C'') > 0$ and $n_{l'}(\mathcal{E}, C'') - n_l(\mathcal{E}, C'') > 1$, because $n_{l'}(\mathcal{E}, C'') > n_{l'}(\mathcal{E}, C') \geq n_l(\mathcal{E}, C') > n_l(\mathcal{E}, C'')$. Therefore, C'' violates the above property, a contradiction.
- $n_l(\mathcal{E}, C') = n_{l'}(\mathcal{E}, C') + 1$: If $d = d' = 1$, this pair of labels does not lead to the distr vectors being different, i.e., there needs to be a different pair of labels for which the number of occurrences differ between C' and C'' as described above (continue the case distinction for a different pair of labels). Otherwise, $n_{l'}(\mathcal{E}, C'') - n_l(\mathcal{E}, C'') > 1$, because $n_{l'}(\mathcal{E}, C'') = n_{l'}(\mathcal{E}, C') + d' = n_l(\mathcal{E}, C') + d' - 1$ and $n_l(\mathcal{E}, C') = n_l(\mathcal{E}, C'') + d$ and $d, d' \geq 1$ and $d + d' > 2$. Therefore, C'' violates the above property, a contradiction.

Thus, each committee satisfying the property has the same diversity according to the indices at hand. In addition, as these diversity indices satisfy *Weak Occurrence Balancing*, each committee $C'' \in \mathcal{R}_{\mathrm{vld}}(\mathcal{E})$ that does not fulfill the property is at most as diverse as C', making C' one of the most diverse committees. □

C.2 Proof of Theorem 2

First, we show the result for LC: In the following, let $\eta = \min\{m, n\}$ and

$$\mu(i, \Gamma) \coloneqq \sum_{j=1}^{i} (\eta + 1))^{n+1-j} \cdot |\sigma_j(\Gamma)| \text{ with } \sigma_j(\Gamma) = \{l \in [m] : n_l(\mathcal{E}, \Gamma) \geq j\}\,,$$
$$L_j(\Gamma) \coloneqq \{l \in [m] : n_l(\mathcal{E}, \Gamma) = j\}$$

and therefore $LC(\mathcal{E}, K) = \mu(k, K)$. Note that for two committees S_1 and S_2 it holds, if $\exists j \in [i]$ with $\mathrm{distr}_j(S_1) \neq \mathrm{distr}_j(S_2)$, that

$$\mu(i, S_1) > \mu(i, K_2) \Leftrightarrow \mathrm{distr}_{m'}(S_1) < \mathrm{distr}_{m'}(S_2)$$

with $m' = \min\{j \in [i] : \mathrm{distr}_j(S_1) \neq \mathrm{distr}_j(S_2)\}$, for an analogous reason to why LC is 1-obvious.

Our polynomial-time algorithm works as follows:

1. Start with a committee K^* which the rule $\mathcal{R}^s$ determines and which therefore has the highest score, set $K = K^*$ and $j = 0$. If $s(\mathcal{E}, K) < \beta$, stop, as there is no committee fulfilling the condition regarding the score.
2. If $j = k$, return K. Otherwise, let

$$I_j = \{i \in [m] : n_i(\mathcal{E}, K) = j \wedge n_i(\mathcal{E}, C) > j\}$$

and, for $i \in I_j$, let $w_i = \max_{c \in C_{label}(\mathcal{E}, C, i) \setminus K} w(\mathcal{E}, c)$, c_i a candidate with label l_i that would contribute the most to score among the candidates with this label that are not part of K, i.e., $c_i \in \{c \in C_{label}(\mathcal{E}, C, i) \setminus K : w(\mathcal{E}, c) = w_i\}$, $X_s = \{c_i : i \in I_j\}$, and $X_e = \{\}$.
3. Pick a $c_a \in \arg\max_{c_i \in X_s \setminus X_e} w_i$. Let

$$C_p = \{c \in K : n_i(\mathcal{E}, K) > j + 1 \text{ with } l_i = \lambda(c)\}.$$

Pick a $c_r \in \arg\min_{c \in C_p} w(\mathcal{E}, c)$, i.e., a candidate which is currently part of K, has a label which occurs more than $j + 1$ times in K, and which contributes the least to the score among such candidates. Let $K' = K \cup \{c_a\} \setminus \{c_r\}$. If $s(\mathcal{E}, K') < \beta$, set $j = j+1$ and go to step 2 (as we cannot increase the number of occurrences of labels occurring j times (further) without decreasing the number of labels occurring at most $j + 1$ times). Otherwise, set $K = K'$, $X_s = X_s \setminus \{c_a\}$ and $X_e = X_e \cup \{c_a\}$. If $|X_s| = 0$ (there are no labels occurring j times left), set $j = j + 1$ and go to step 2. Otherwise, start with step 3 again.

We show by induction that, when visiting the second step for the i-th time, with $i \in \{2, \dots k + 1\}$ and K_i being the committee when reaching this step for the i-th time, there is

1. no other committee $K' \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ with $s(\mathcal{E}, K') \geq \beta$ and $\mu(i-1, K') > \mu(i-1, K_i)$ (later referred to as the first condition).
2. no other committee $K'' \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ with $\mu(i-1, K'') = \mu(i-1, K_i)$ and $s(\mathcal{E}, K'') > s(\mathcal{E}, K_i)$ (later referred to as the second condition).

Note that these conditions imply the following (later referred to as the third condition):

$$\forall l \in [m], l' \in \{\phi \in [m] : n_\phi(\mathcal{E}, K_i) < n_l(\mathcal{E}, K_i)\},$$
$$c \in C_{label}(\mathcal{E}, K_i, l), c' \in C_{label}\big(\mathcal{E}, C \setminus K_i, l'\big) : w\big(\mathcal{E}, c'\big) \leq w(\mathcal{E}, c).$$

Assume this is not the case, i.e., there are l, l', c, c' so that $w(\mathcal{E}, c') > w(\mathcal{E}, c)$. Let $K_d = K_i \setminus \{c\} \cup \{c'\}$. Therefore, $s(\mathcal{E}, K_d) > s(\mathcal{E}, K_i) \geq \beta$. In addition, if $n_{l'}(\mathcal{E}, K_i) \geq i - 1$ or $n_l(\mathcal{E}, K_i) = n_{l'}(\mathcal{E}, K_i) + 1$, it follows that $\mu(i-1, K_d) = \mu(i-1, K_i)$, which contradicts the second condition. On the other hand, if $n_{l'}(\mathcal{E}, K_i) < i - 1$ and $n_l(\mathcal{E}, K_i) > n_{l'}(\mathcal{E}, K_i) + 1$, it follows that $\mu(i-1, K_d) > \mu(i-1, K_i)$, which contradicts the first condition.

We start with $i = 2$, i.e., with visiting the second step for the second time: If $K_1 = K^*$ has either no label occurring 0 times, or l_p labels occurring 0 times and all other committees of size k which satisfy the bound β have at least l_p labels occurring 0 times, $s(\mathcal{E}, K_1) = s(\mathcal{E}, K_2) \geq s(\mathcal{E}, K')$ for all other committees K' of size k, as $K_1 = K^* = K_2$. Otherwise, let $K' \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ be a committee respecting β in which $l_n < l_p$ labels occur 0 times and, therefore, in which more labels occur at least once. We show that K_2 has at most l_n labels occurring 0 times and, if it has exactly l_n many labels occurring 0 times, a score which is at least as good.

First, we transform K' so that the score does not decrease and the number of labels occurring 0 times does not increase. As long as $\exists l' \in [m] : n_{l'}(\mathcal{E}, K') = 0 < n_{l'}(\mathcal{E}, K_1)$ and therefore $\exists l \in [m] : n_l(\mathcal{E}, K_1) = 0 < n_l(\mathcal{E}, K')$, we set $K' = K' \setminus \{c\} \cup \{c'\}$ with $c' \in C_{label}(\mathcal{E}, K_1, l')$ and $c \in C_{label}(\mathcal{E}, K', l)$, which does not reduce the score (as K_1 maximizes the score). Let l_n be the number of labels occurring 0 times in this updated K'.

Pick the $L_a \subseteq L_0(K_1) \setminus L_0(K')$ with $|L_a| = l_p - l_n$ and for each $l \in L_a$ a candidate $c_a^{(l)} \in C_{label}(\mathcal{E}, K', l)$ and let $C_a = \bigcup_{l \in L_a} \left\{c_a^{(l)}\right\}$. There needs to be a $C_r \subseteq C$ with $|C_r| = |C_a| = l_p - l_n$ such that

$\forall c \in C_r : c \in K_1 \setminus K'$ and $n_\theta(\mathcal{E}, K_1) - n_\theta(\mathcal{E}, C_r) \geq n_\theta(\mathcal{E}, K') > 0$ with $l_\theta = \lambda(c)$, as it needs to hold that $k = \sum_{i\in[m]} n_i(K_1) = \sum_{i\in[m]\setminus L_0(K_1)} n_i(K_1) = \sum_{i\in[m]\setminus L_0(K_1)} n_i(K') + \sum_{i\in L_a} n_i(K')$ with $\sum_{i\in L_a} n_i(K') \geq l_p - l_n$.

It holds that $s(\mathcal{E}, K' \cup C_r \setminus C_a) \leq s(\mathcal{E}, K_1)$ and therefore it holds that $s(\mathcal{E}, K_1 \setminus C_r \cup C_a) \geq s(\mathcal{E}, K') \geq \beta$. Consequently, the algorithm will repeat the third step at least $l_p - l_n$ times, in each step removing a candidate from C_r or, alternatively, a candidate c' of a label occurring at least 2 times with a lower $w(\mathcal{E}, c')$, and in each step adding a candidate from C_a or a candidate c'' of a label occurring 0 times with a larger $w(\mathcal{E}, c'')$. Thus, K_2 has at most l_n labels occurring 0 times and, if exactly l_n labels occur 0 times, a score at least as good as the score of K'.

Now, the inductive step $i \rightsquigarrow i+1$ follows: Let l_p be the number of labels occurring $i-1$ times in K_i. Note that for a committee $K' \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ with $\mu(i, K') \geq \mu(i, K_i)$ and $\text{score}_{\text{AV}}(K') \geq \beta$, it has to hold that $\mu(i-1, K') = \mu(i-1, K_i)$ and thus $\sigma_j(K') = \sigma_j(K_i)$ for $j \in [i-1]$ and $|L_j(K')| = |L_j(K_i)|$ for $j \in \{0, \ldots, i-2\}$.

If $l_p = 0$ (and therefore the number of labels occurring at least i times is optimal) or all other committees $K' \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ which satisfy the bound β and for which $\mu(i-1, K') = \mu(i-1, K_i)$ have at least l_p many labels occurring $i-1$ times (and thus at most as many labels occurring at least i times as K_i), $K_{i+1} = K_i$ and $s(\mathcal{E}, K_i) \geq s(\mathcal{E}, K')$ due to the induction hypothesis.

Otherwise, let $K' \in \mathcal{R}_{\text{vld}}(\mathcal{E})$ be a committee respecting β in which $l_n < l_p$ labels occur $i-1$ times and $\mu(i-1, K') = \mu(i-1, K_i)$ and thus more labels occur at least i times. We show that K_{i+1} has at most l_n labels occurring $i-1$ times and, if exactly l_n labels occur $i-1$ times, a score which is at least as good.

First, we transform K' so that $\mu(i-1, K')$ remains unchanged, the score does not decrease, and the number of labels occurring exactly $i-1$ times does not increase. First, note that it is not possible that $\exists l, l' \in [m] : n_l(\mathcal{E}, K') \leq i-2$, $n_l(\mathcal{E}, K') = n_{l'}(\mathcal{E}, K_i) < n_l(\mathcal{E}, K_i)$, and $n_{l'}(\mathcal{E}, K') \geq n_l(\mathcal{E}, K') + 2$: If this were possible, $s(\mathcal{E}, K' \setminus \{c'\} \cup \{c\}) \geq s(\mathcal{E}, K') \geq \beta$ with $c' \in C_{label}(\mathcal{E}, K', l') \setminus K_i, c \in C_{label}(\mathcal{E}, K_i, l) \setminus K'$ because of the third condition and, therefore, $\mu(i-1, K' \setminus \{c'\} \cup \{c\}) > \mu(i-1, K_i)$, a contradiction to the induction hypothesis.

For $j \in (0, \ldots, i-2)$, we do the following: As long as $\exists l \in [m] : n_l(\mathcal{E}, K') = j < n_l(\mathcal{E}, K_i)$ and therefore $\exists l' \in [m] : n_{l'}(\mathcal{E}, K_i) = j < n_{l'}(\mathcal{E}, K') = j+1$ (see the previous note), we set $K' = K' \setminus \{c'\} \cup \{c\}$ with $c' \in C_{label}(\mathcal{E}, K', l') \setminus K_i$ and $c \in C_{label}(\mathcal{E}, K_i, l) \setminus K'$, which leaves $\mu(i-1, K')$ unchanged and does not reduce the score because of the third condition. This results in

$$\begin{aligned} L_d &\coloneqq \left\{l \in [m] : n_l(\mathcal{E}, K_i) = n_l\big(\mathcal{E}, K'\big) \leq i-2\right\} \\ &= \left\{l \in [m] : n_l(\mathcal{E}, K_i) \leq i-2 \vee n_l\big(\mathcal{E}, K'\big) \leq i-2\right\}. \end{aligned}$$

Then, for $j = i-1$, we do the following: As long as

- $\exists l' \in [m] : n_{l'}(\mathcal{E}, K_i) > n_{l'}(\mathcal{E}, K') = j$
- and therefore $\exists l \in [m] : n_l(\mathcal{E}, K') > n_l(\mathcal{E}, K_i) = j$ (because $|L_j(K_i)| > |L_j(K')|$)
- and $\exists C_a \subseteq K_i \setminus K' : |C_a| = n_l(\mathcal{E}, K') - n_l(\mathcal{E}, K_i)$ such that $\forall c \in C_a : n_\theta(\mathcal{E}, K') + n_\theta(\mathcal{E}, C_a) \leq n_\theta(\mathcal{E}, K_i)$ with $l_\theta = \lambda(c)$, and $\exists c_{l'} \in C_a : c_{l'} \in C_{label}(\mathcal{E}, K_i, l')$ (such a C_a exists, as it holds that $\sum_{\phi\in[m]\setminus L_d\setminus\{l\}} n_\phi(K_i) + n_l(\mathcal{E}, K_i) = \sum_{\phi\in[m]\setminus L_d\setminus\{l\}} n_\phi(K') + n_l(\mathcal{E}, K')$ and $n_l(\mathcal{E}, K_i) < n_l(\mathcal{E}, K')$),

set $K' = K' \setminus C_r \cup C_a$ with $C_r \subseteq C_{label}(\mathcal{E}, K', l) \setminus K_i$ and $|C_r| = |C_a|$ so that l' occurs more often than j times in K', but now l occurs j times. This does not reduce the score either because of the third condition, as we decrease the number of occurrences of label l' in K' (by removing candidates with this label that are not part of K_i) and as we increase the number of labels which occur more than j times in K_i (by adding candidates with this label that are part of K_i which have a higher score s than the removed candidates because of the third condition).

For the resulting K', let l_n be the number of labels occurring $i-1$ times in this updated K', for which $\forall l \in L_{i-1}(K') : n_l(\mathcal{E}, K') = n_l(\mathcal{E}, K_i)$. Next, take $L_a \subseteq L_{i-1}(K_i) \setminus L_{i-1}(K')$ for which it holds that $|L_a| = l_p - l_n$ such that $\forall l_a \in L_a : C_{label}(\mathcal{E}, K', l_a) > i-1$, pick a $c_{l_a} \in C_{label}(\mathcal{E}, K', l_a) \setminus K_i$ for all $l_a \in L_a$, and let $C_a = \bigcup_{l_a \in L_a} c_{l_a}$. Additionally, there needs to be a $C_r \subseteq K_i \setminus K'$ with $|C_r| = |C_a|$ such that $\forall c \in C_r : n_\theta(\mathcal{E}, K_i) > n_\theta(\mathcal{E}, K') > i-1 \wedge n_\theta(\mathcal{E}, K_i) - n_\theta(\mathcal{E}, C_r) \geq n_\theta(\mathcal{E}, K')$ with $l_\theta = \lambda(c)$: Such a C_r exists, as it holds, with $L_i^{\geq} = [m] \setminus L_d \setminus L_{i-1}(K_i)$ as the indices of labels which occur at least i times in K_i, that

$$\begin{aligned}
&\sum_{l \in L_i^{\geq}} n_l(\mathcal{E}, K_i) + \sum_{L_{i-1}(K_i)} n_l(\mathcal{E}, K_i) = \sum_{l \in L_i^{\geq}} n_l\big(\mathcal{E}, K'\big) + \sum_{L_{i-1}(K_i)} n_l\big(\mathcal{E}, K'\big) \\
\Leftrightarrow &\sum_{l \in L_i^{\geq}} n_l(\mathcal{E}, K_i) = \sum_{l \in L_i^{\geq}} n_l\big(\mathcal{E}, K'\big) + \sum_{L_{i-1}(K_i)} n_l\big(\mathcal{E}, K'\big) - l_p\,(i-1) \\
&\geq \sum_{l \in L_i^{\geq}} n_l\big(\mathcal{E}, K'\big) + l_p - l_n.
\end{aligned}$$

The last inequality holds because $\sum_{l \in L_{i-1}(K_i)} n_l(\mathcal{E}, K') \geq l_n\,(i-1) + (l_p - l_n)\,i$ and $l_n\,(i-1) + (l_p - l_n)\,i - l_p\,(i-1) = (l_p - l_n)\,i - (l_p - l_n)\,(i-1) = l_p - l_n$.

$s(\mathcal{E}, K' \cup C_r \setminus C_a) \leq s(\mathcal{E}, K_i)$ holds (because of the induction hypothesis and the third condition) and hence $s(\mathcal{E}, K_i \setminus C_r \cup C_a) \geq s(\mathcal{E}, K') \geq \beta$. Consequently, the algorithm will repeat the third step at least $l_p - l_n$ times, in each step removing a candidate from C_r or, alternatively, a candidate c' of a label occurring more than i times with a lower $w(\mathcal{E}, c')$, and in each step adding a candidate from C_a or a candidate c'' of a label occurring $i-1$ times with a larger $w(\mathcal{E}, c'')$. Thus, K_{i+1} has at most l_n labels occurring $i-1$ times and, if exactly l_n labels occur $i-1$ times, a score at least as good as the score of K'.

The result for Ri follows directly from Corollary 3, which states that each optimal solution of LC is also an optimal solution of Ri. □

C.3 Proof of Corollary 1

For score_{AV} it holds that $w(\mathcal{E}, c) = |\,\{a \in A : c \in U(a)\}\,| \leq |A|$, for each $c \in C$.

For Si, it holds that maximizing $-\sum_{i \in [m]} p_i^2$ yields the same solutions (with a different objective value) as maximizing $\left(\sum_{i \in [m]} -n_i^2\right) + 2k^2 = \sum_{i \in [m]} -n_i^2 + 2 \cdot n_i \cdot k$. Thus, $t(i) = -2i + 1 + 2k$ can be chosen (which is strictly monotonically decreasing), as it holds that $1 \leq t(i) \leq 2k+1$ for $i \in [k]$ and $\sum_{i=1}^{n} 2k - 2i + 1 = 2kn - n^2$.

For Sh, let $M_r = \{i \in [m] : p_i(\mathcal{E}, S) > 0\}$. It holds that maximizing

$$\begin{aligned}
-\sum_{i \in M_r} p_i \cdot \log(p_i) &= -\sum_{i \in M_r} \frac{n_i}{k} \cdot \log\left(\frac{n_i}{k}\right) = -\frac{1}{k}\left(\sum_{i \in M_r} n_i \cdot (\log(n_i) - \log(k))\right) \\
&= -\frac{1}{k}\left(-k \cdot \log(k) + \sum_{i \in M_r} n_i \cdot \log(n_i)\right)
\end{aligned}$$

yields the same results as maximizing $-\sum_{i \in M_r} n_i \cdot \log(n_i)$ which, in turn, yields the same results as maximizing

$$k \cdot (\log(k) + 2) - \sum_{i \in M_r} n_i \cdot \log(n_i) = -\sum_{i \in M_r} n_i\,(\log(n_i) - \log(k) - 2)$$

Thus, $t(i) = -i\log(i) + (i-1)\log(i-1) + \log(k) + 2$ with $t(1) = \log(k) + 2$ can be chosen, which is strictly monotonically decreasing, and $t(i) > 0$ for $i \in [k]$:

Let $f(x) = x\log(x)$ with $f'(x) = 1 + \log(x)$. According to the mean value theorem, $\exists c \in (i-1, i)$ such that $f(i) - f(i-1) = f'(c) = 1 + \log(c)$. Thus, it holds for all $i \in [k]$ that

$$t(i) = -i\log(i) + (i-1)\log(i-1) + \log(k) + 2 = -1 - \log(c) + \log(k) + 2 > 0,$$

as $\log(c) \leq \log(k)$ for $i \in [k]$. □

C.4 Proof of Theorem 4

Given an instance I_D of (D, s)-DSCR with n candidates, β as the score bound, and δ as the diversity bound, we construct an instance I_K of the 0-1 Knapsack problem in polynomial time. We assume that $\beta \geq 0$; otherwise, we only need to optimize the diversity, which is in $\mathbf{P}$ for such diversity indices, as the same approach as in the proof of Observation 7 can be used because $t(i)$ is strictly monotonically decreasing. In addition, we assume that $\delta \geq 0$; otherwise, the problem is in $\mathbf{P}$, as we only need to find a committee of size k fulfilling the score constraint which, if the problem is feasible, can be achieved by choosing k many candidates with the highest weights. Furthermore, we assume that $\delta \leq k\zeta$ (which can be checked easily), as the problem is infeasible otherwise.

For each candidate c_i, we add an item x_i with the weight

$$w_K(x_i) \coloneqq -t(\pi(c_i)) + \eta \text{ with } \eta \coloneqq k\zeta + \zeta + 1,$$

where π outputs c_i's position in a descending ordering of the candidates with the same label as c_i based on w, and the value

$$v(x_i) \coloneqq w(c_i) + n\alpha + 1 \text{ where } \alpha \coloneqq \max_{c \in C} w(\mathcal{E}, c).$$

Thus, between two candidates c_i, c_j with the same label of which c_i contributes more to the score, i.e., $w(c_i) > w(c_j)$, it holds that $v(x_i) > v(x_j)$ and $t(\pi(x_i)) > t(\pi(x_j))$ since $\pi(x_i) < \pi(x_j)$, and hence $w_K(x_i) < w_K(x_j)$. Therefore, replacing an item x_j in a solution to I_K by an item x_i with the same label but with a smaller value of π will lead to a solution to I_K with a higher value and a lower weight. Furthermore, we set the knapsack's bound to

$$B \coloneqq -\delta + k\eta.$$

Let, for a solution X of I_K, $S(X) = \{c_i \mid x_i \in X\}$, $v(X)$ and $w_K(X)$ be the value and weight of X, and, for a solution S of I_D, $X(S) = \{x_i \mid c_i \in S\}$. We claim the following:

1. If X is a solution to I_K with value at least $u_v \coloneqq \beta + k\,(n\alpha + 1)$, then $S(X)$ is a solution to I_D.
2. If S is a solution to I_D, then there is a (possibly different) solution S' to I_D such that $D(S) = D(S')$ and $X(S')$ is a solution to I_K with value at least u_v.

1. Let X be a solution to I_K with value at least u_v. It follows that $|X| \geq k$, otherwise

$$v(X) \leq (k-1)\,(n\alpha + 1 + \alpha) = k\,(n\alpha + 1) + k\alpha - n\alpha - 1 - \alpha \overset{k \leq n}{<} k\,(n\alpha + 1) \leq u_v,$$

a contradiction. Next, assume that $|X| = k + i > k$, then

$$w_K(X) \geq (k+1)\,(-\zeta + \eta) = (k+1)\,(k\zeta + 1) = k^2\zeta + k + k\zeta + 1 > k^2\zeta + k\zeta + k = k\eta \geq B,$$

a contradiction. Thus, $|X| = k$. In addition, $-\delta + k\eta = B \geq w_K(X) \geq k\eta - D(S(X)) \Leftrightarrow \delta \leq D(S(X))$ and $v(X) = k\,(n\alpha + 1) + w(S(X)) \geq u_v = \beta + k\,(n\alpha + 1) \Leftrightarrow w(S(X)) \geq \beta$. Thus, $S(X)$ fulfills the score and diversity constraints and is therefore a solution to I_D.

2. Let S_D^* be a solution to I_D. Consider the committee S^* with $n_l(\mathcal{E}, S^*) = n_l(\mathcal{E}, S_D^*)$ for every $l \in [m]$ and $c \in S^* \Leftrightarrow \pi(c) \leq n_j(\mathcal{E}, S_D^*)$ with $l_j = \lambda(c)$. Thus,

$$(D(\mathcal{E}, S^*) = -w(X(S^*)) + k\eta = D(\mathcal{E}, S_D^*) \geq \delta) \Rightarrow (w(X(S^*)) \leq -\delta + k\eta)$$
$$(w(S^*) \geq w(S_D^*) \geq \beta) \Rightarrow (v(X(S^*)) = w(S^*) + k\,(n\alpha + 1) \geq \beta + k\,(n\alpha + 1) = u_v)\,.$$

Therefore, $X(S^*)$ is a solution to I_K.

Thus, the problem can be solved by the solving the 0-1 Knapsack instance with dynamic programming in $\mathcal{O}(nB) = \mathcal{O}(n\,(k\eta - \delta))$ time. □

D Additional Material for Section 5.2

The dimension of the experimental data when using $k = 8$ and $k = 6$ can be seen in Fig. 3. Note, that the number of instances increases for smaller k (729 instances for $k = 8$, 773 for $k = 6$), because we discard instances with at most k many candidates.

For each diversity index considered, the proportion of the optimal diversity reached over all experimental data for the different rules (including the rules that represent our approaches of incorporating diversity, i.e., $\mathcal{R}^{-1}_{\text{sat}}$ and $\mathcal{R}^{p}_{\text{scr}}$) are visualized using box plots in Figs. 4 to 15, which also include results for $k \in \{6, 8, 10\}$ and seq-Phragmén, Rule X, PAV, and CC.

Table 1 shows the average proportion of the optimal diversity that a rule reaches for each $k \in \{6, 8, 10\}$, diversity index, and rule considered. Similarly, the number of instances for which the rule achieves the optimal diversity can be seen in Table 2. These results support the qualitative results mentioned in the main body of the paper (including the footnotes). Here, we want to highlight the following, additional observations based on these tables and plots:

As stated in the main body of the paper, a higher proportion of the optimal diversity is reached on average when using $\mathcal{R}^{90}_{\text{scr}}$ than when using $\mathcal{R}^{-1}_{\text{sat}}$ with $\mathcal{R} \in \{\text{AV}, \text{SAV}\}$, with only a few exceptions for $k = 6$: These two proportions are the same for Ri together with AV, Sh together with SAV, and LC together with SAV; a higher proportion of the optimal diversity is reached on average when using $\mathcal{R}^{-1}_{\text{sat}}$ than when using $\mathcal{R}^{90}_{\text{scr}}$ (i.e., the other way around) if SAV is used together with Ri or Si.

When decreasing k, it holds that, for the k investigated and for each rule and diversity index considered, the proportion of instances for which the optimal diversity is reached increases, with the only exceptions occurring when looking at $\text{AV}^{40}_{\text{scr}}$, $\text{AV}^{30}_{\text{scr}}$, $\text{SAV}^{50}_{\text{scr}}$, $\text{AV}^{40}_{\text{scr}}$, and $\text{AV}^{30}_{\text{scr}}$ (i.e. when the scoring constraints are relatively weak), for which the proportions stay the same and the optimal diversity is already reached for at least 97% of the instances with $k = 10$. The average percentage of the optimal diversity reached also increases in most cases when decreasing k and $\mathcal{R}^{-1}_{\text{sat}}$ is considered, with only a few exceptions when using Sh. However, there are far more exceptions in which this percentage stays the same when looking at $\mathcal{R}^{p}_{\text{scr}}$.

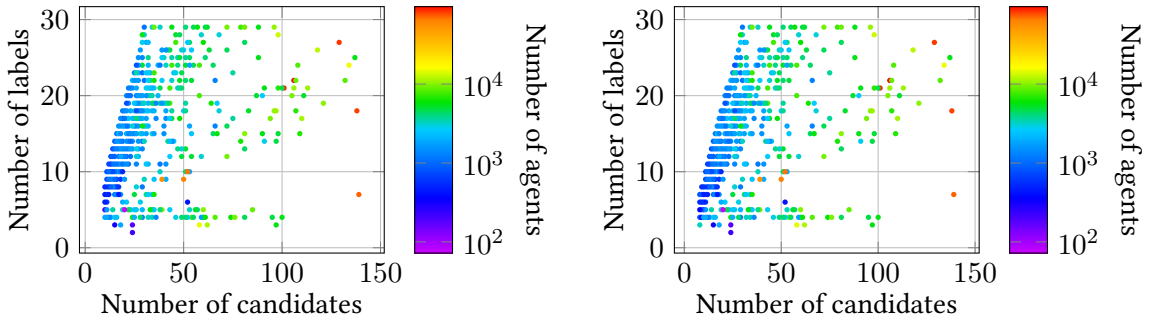
Lastly, we want to highlight that the optimal diversity is reached for at least 73% of the data when reducing the score to be achieved by 20% of the optimal score (i.e. $\mathcal{R}^{80}_{\text{scr}}$) for each combination of diversity index, k, and rule considered, and for at least 84% of the data when allowing a reduction of the score by 30%.

Index	Ri			Sh			Si			LC		
k	10	8	6	10	8	6	10	8	6	10	8	6
AV	76	78	80	80	81	81	66	69	74	77	79	82
$\text{AV}^{-1}_{\text{sat}}$	86	89	92	88	90	92	78	83	88	86	89	92
$\text{AV}^{90}_{\text{scr}}$	91	92	92	93	93	93	85	87	89	92	92	93
$\text{AV}^{80}_{\text{scr}}$	96	96	96	96	97	97	92	93	95	96	96	97
$\text{AV}^{70}_{\text{scr}}$	98	98	98	98	98	98	96	97	97	98	98	98
$\text{AV}^{60}_{\text{scr}}$	99	99	99	99	99	99	98	98	99	99	99	99
$\text{AV}^{50}_{\text{scr}}$	100	100	100	100	100	100	99	100	100	100	100	100
$\text{AV}^{40}_{\text{scr}}$	100	100	100	100	100	100	100	100	100	100	100	100
$\text{AV}^{30}_{\text{scr}}$	100	100	100	100	100	100	100	100	100	100	100	100
SAV	77	79	81	81	82	83	67	70	76	78	80	83
$\text{SAV}^{-1}_{\text{sat}}$	87	90	93	89	91	93	80	84	90	87	90	93
$\text{SAV}^{90}_{\text{scr}}$	92	92	92	93	93	93	86	87	89	92	92	93
$\text{SAV}^{80}_{\text{scr}}$	96	96	96	97	97	97	93	94	95	96	97	97
$\text{SAV}^{70}_{\text{scr}}$	98	98	98	98	98	98	96	96	97	98	98	98
$\text{SAV}^{60}_{\text{scr}}$	99	99	99	99	99	99	98	98	99	99	99	99
$\text{SAV}^{50}_{\text{scr}}$	100	100	100	100	100	100	99	99	100	100	100	100
$\text{SAV}^{40}_{\text{scr}}$	100	100	100	100	100	100	100	100	100	100	100	100
$\text{SAV}^{30}_{\text{scr}}$	100	100	100	100	100	100	100	100	100	100	100	100
PAV	78	79	82	82	82	83	67	71	76	78	80	83
$\text{PAV}^{-1}_{\text{sat}}$	87	90	93	89	91	93	80	84	90	88	90	93
seq-Phragmén	77	79	82	81	82	83	67	71	76	78	80	84
seq-$\text{Phragmén}^{-1}_{\text{sat}}$	87	90	93	89	91	93	80	84	90	87	90	93
Rule X	77	79	82	81	82	83	67	71	76	78	81	83
Rule $\text{X}^{-1}_{\text{sat}}$	87	90	93	89	91	93	79	85	90	87	90	93
CC	81	83	84	84	85	85	71	75	79	82	84	85
Rule $\text{CC}^{-1}_{\text{sat}}$	90	92	94	92	93	95	84	88	92	91	93	95

Table 1: For each diversity index in $\{Ri, Sh, Si\}$, each rule considered, and each $k \in \{6, 8, 10\}$, the average percentage of the optimal diversity that the rule reaches is stated.

Index	Ri			Sh			Si			LC		
k	10	8	6	10	8	6	10	8	6	10	8	6
AV	17	23	35	14	22	35	14	22	35	14	22	35
$\text{AV}^{-1}_{\text{sat}}$	42	55	69	36	52	68	36	52	68	36	52	68
$\text{AV}^{90}_{\text{scr}}$	59	64	71	54	60	70	54	60	70	54	60	70
$\text{AV}^{80}_{\text{scr}}$	78	81	84	73	77	83	73	77	83	73	77	83
$\text{AV}^{70}_{\text{scr}}$	88	90	92	84	87	91	84	87	91	84	87	91
$\text{AV}^{60}_{\text{scr}}$	92	94	96	91	92	95	91	92	95	91	92	95
$\text{AV}^{50}_{\text{scr}}$	97	98	99	97	98	99	97	98	99	97	98	99
$\text{AV}^{40}_{\text{scr}}$	99	100	100	99	100	100	99	100	100	99	100	100
$\text{AV}^{30}_{\text{scr}}$	100	100	100	100	100	100	100	100	100	100	100	100
SAV	17	23	36	13	21	36	13	21	36	13	21	36
$\text{SAV}^{-1}_{\text{sat}}$	43	58	75	38	54	73	38	54	73	38	54	73
$\text{SAV}^{90}_{\text{scr}}$	57	63	70	51	59	69	51	59	69	51	59	69
$\text{SAV}^{80}_{\text{scr}}$	79	83	84	74	79	83	74	79	83	74	79	83
$\text{SAV}^{70}_{\text{scr}}$	88	89	91	85	86	90	85	86	90	85	86	90
$\text{SAV}^{60}_{\text{scr}}$	92	94	95	90	92	95	90	92	95	90	92	95
$\text{SAV}^{50}_{\text{scr}}$	97	97	99	97	97	99	97	97	99	97	97	99
$\text{SAV}^{40}_{\text{scr}}$	99	100	100	99	100	100	99	100	100	99	100	100
$\text{SAV}^{30}_{\text{scr}}$	100	100	100	100	100	100	100	100	100	100	100	100
PAV	17	24	38	14	22	37	14	22	37	14	22	37
$\text{PAV}^{-1}_{\text{sat}}$	44	59	73	38	55	72	38	55	72	38	55	72
seq-Phragmén	18	24	39	14	22	38	14	22	38	14	22	38
$\text{seq-Phragmén}^{-1}_{\text{sat}}$	43	58	74	38	54	72	38	54	72	38	54	72
Rule X	17	24	38	13	22	38	13	22	38	13	22	38
$\text{Rule X}^{-1}_{\text{sat}}$	44	59	74	38	54	72	38	54	72	38	54	72
CC	23	32	42	18	28	41	18	28	41	18	28	41
$\text{Rule CC}^{-1}_{\text{sat}}$	56	69	81	52	66	79	52	66	79	52	66	79

Table 2: For each diversity index in $\{Ri, Sh, Si\}$, each rule considered, and each $k \in \{6, 8, 10\}$, the percentage of instances for which the rule achieved the optimal diversity is stated.



(a) $k = 8$.

(b) $k = 6$.

Figure 3: The dimensions of the experimental data, where the color of each point represents the average number of agents of all instances with the given number of labels and candidates.

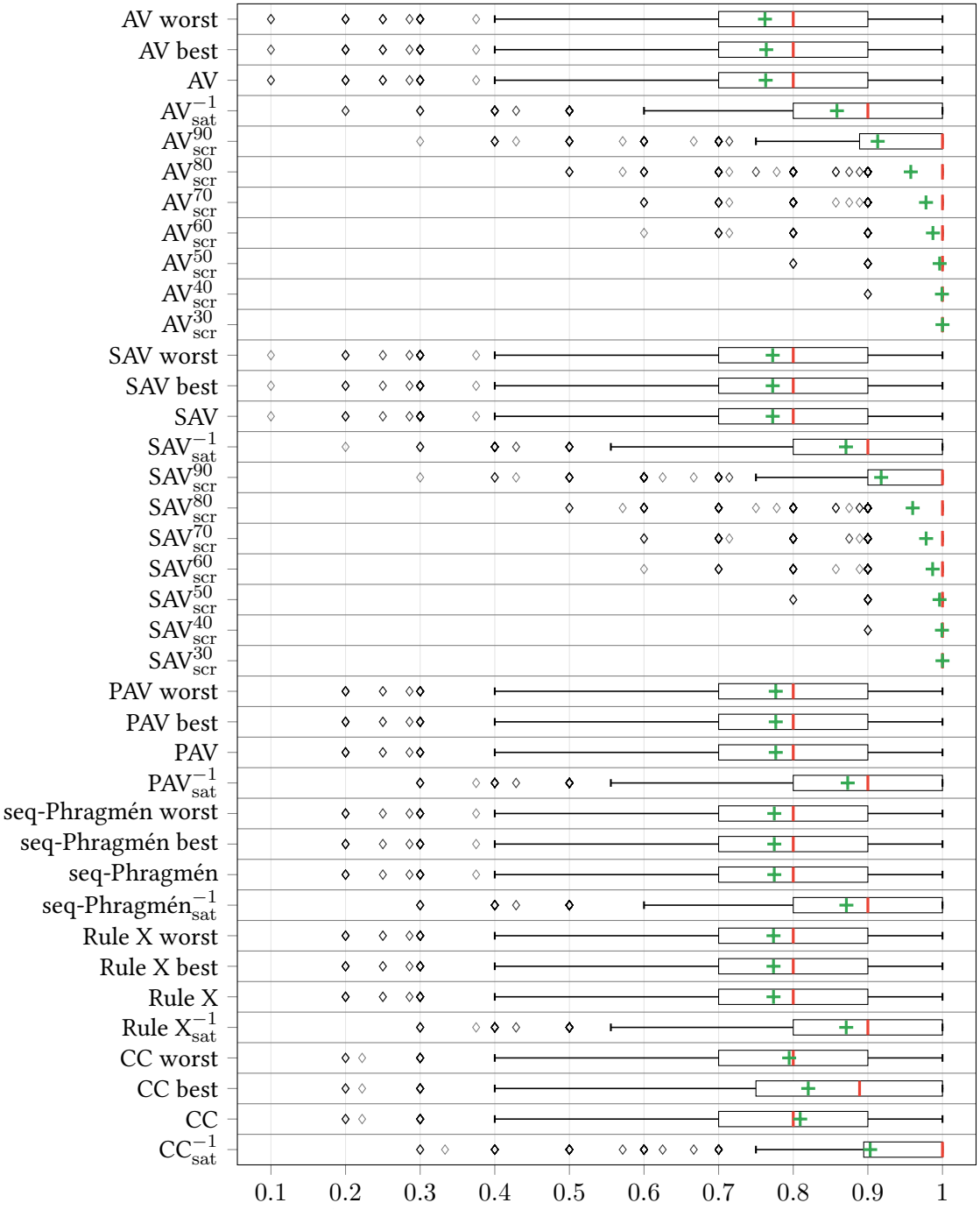


Figure 4: The proportion of the optimal diversity reached over all experimental data with $k = 10$ when using Ri for the different rules and approaches we consider. "$\mathcal{R}$ best" ("$\mathcal{R}$ worst") refers to the rule that chooses the committees with the highest (lowest) diversities among the winning committees of $\mathcal{R}$. If only $\mathcal{R}$ is written, the diversity of the winning committee that *abcvoting* returns for $\mathcal{R}$ is considered. The red line indicates the median, the green cross the mean.

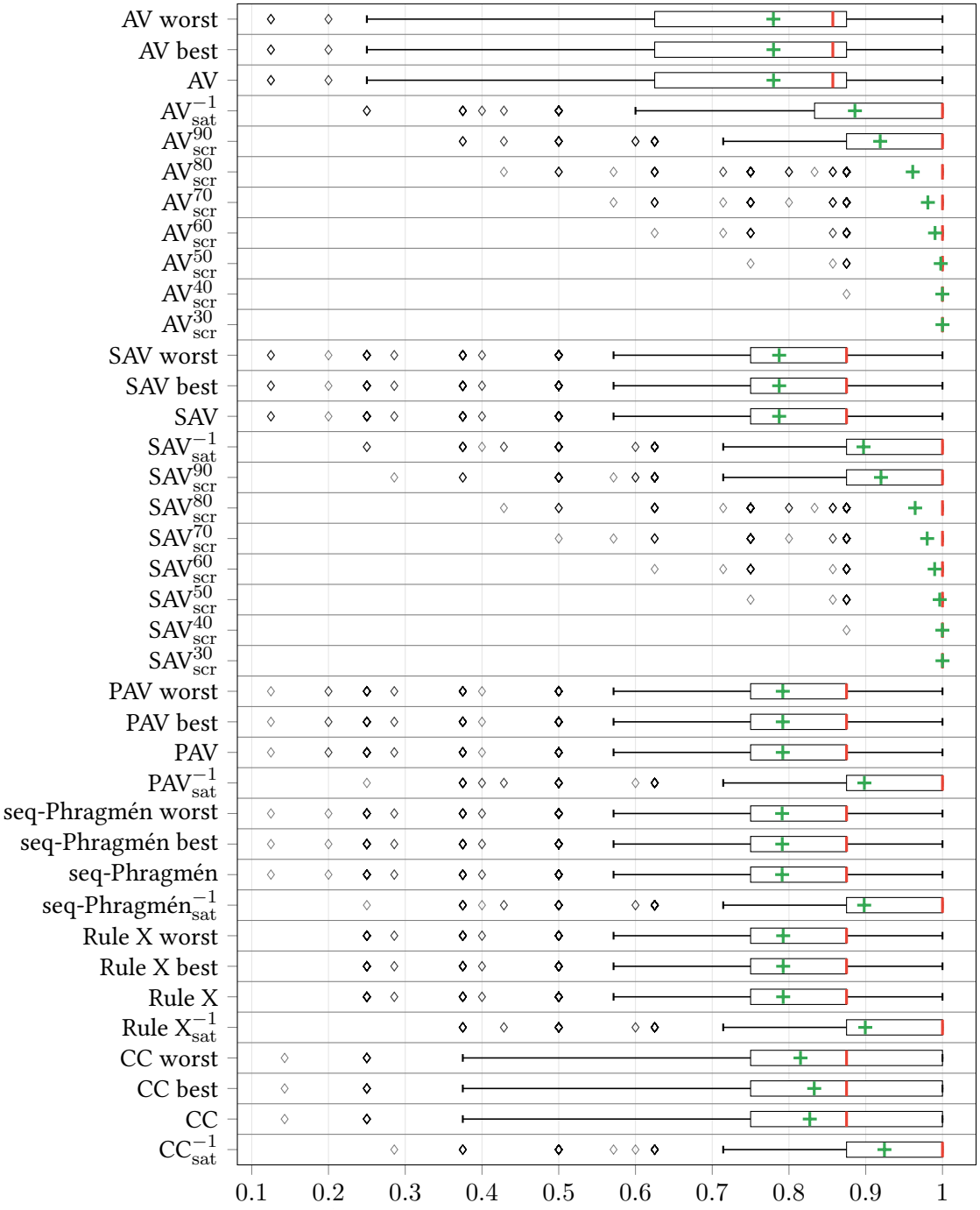


Figure 5: The proportion of the optimal diversity reached over all experimental data with $k = 8$ when using Ri for the different rules and approaches we consider. "$\mathcal{R}$ best" ("$\mathcal{R}$ worst") refers to the rule that chooses the committees with the highest (lowest) diversities among the winning committees of $\mathcal{R}$. If only $\mathcal{R}$ is written, the diversity of the winning committee that *abcvoting* returns for $\mathcal{R}$ is considered. The red line indicates the median, the green cross the mean.

Figure 6: The proportion of the optimal diversity reached over all experimental data with $k = 6$ when using Ri for the different rules and approaches we consider. "$\mathcal{R}$ best" ("$\mathcal{R}$ worst") refers to the rule that chooses the committees with the highest (lowest) diversities among the winning committees of $\mathcal{R}$. If only $\mathcal{R}$ is written, the diversity of the winning committee that *abcvoting* returns for $\mathcal{R}$ is considered. The red line indicates the median, the green cross the mean.

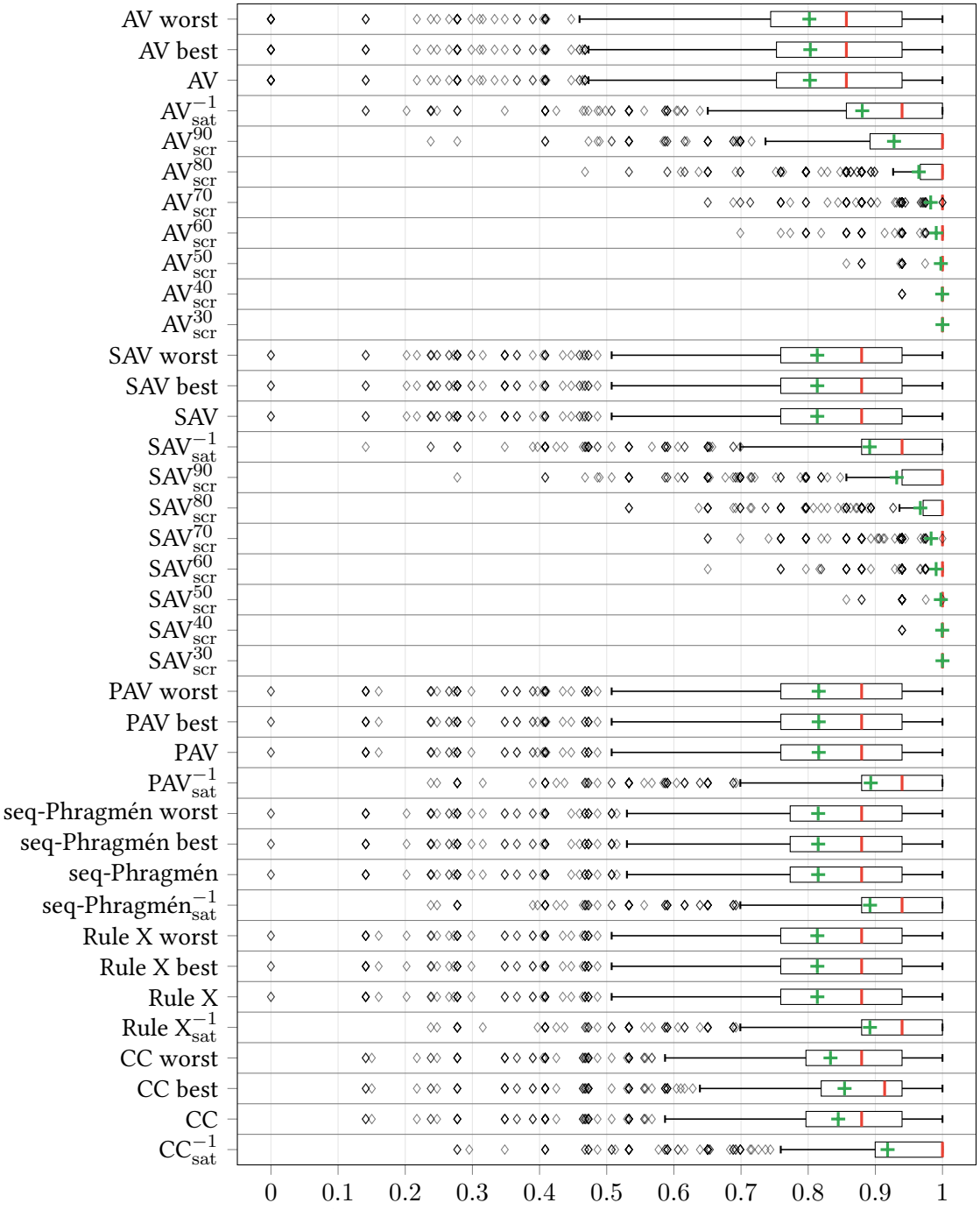


Figure 7: The proportion of the optimal diversity reached over all experimental data with $k = 10$ when using *Sh* for the different rules and approaches we consider. "$\mathcal{R}$ best" ("$\mathcal{R}$ worst") refers to the rule that chooses the committees with the highest (lowest) diversities among the winning committees of $\mathcal{R}$. If only $\mathcal{R}$ is written, the diversity of the winning committee that *abcvoting* returns for $\mathcal{R}$ is considered. The red line indicates the median, the green cross the mean.

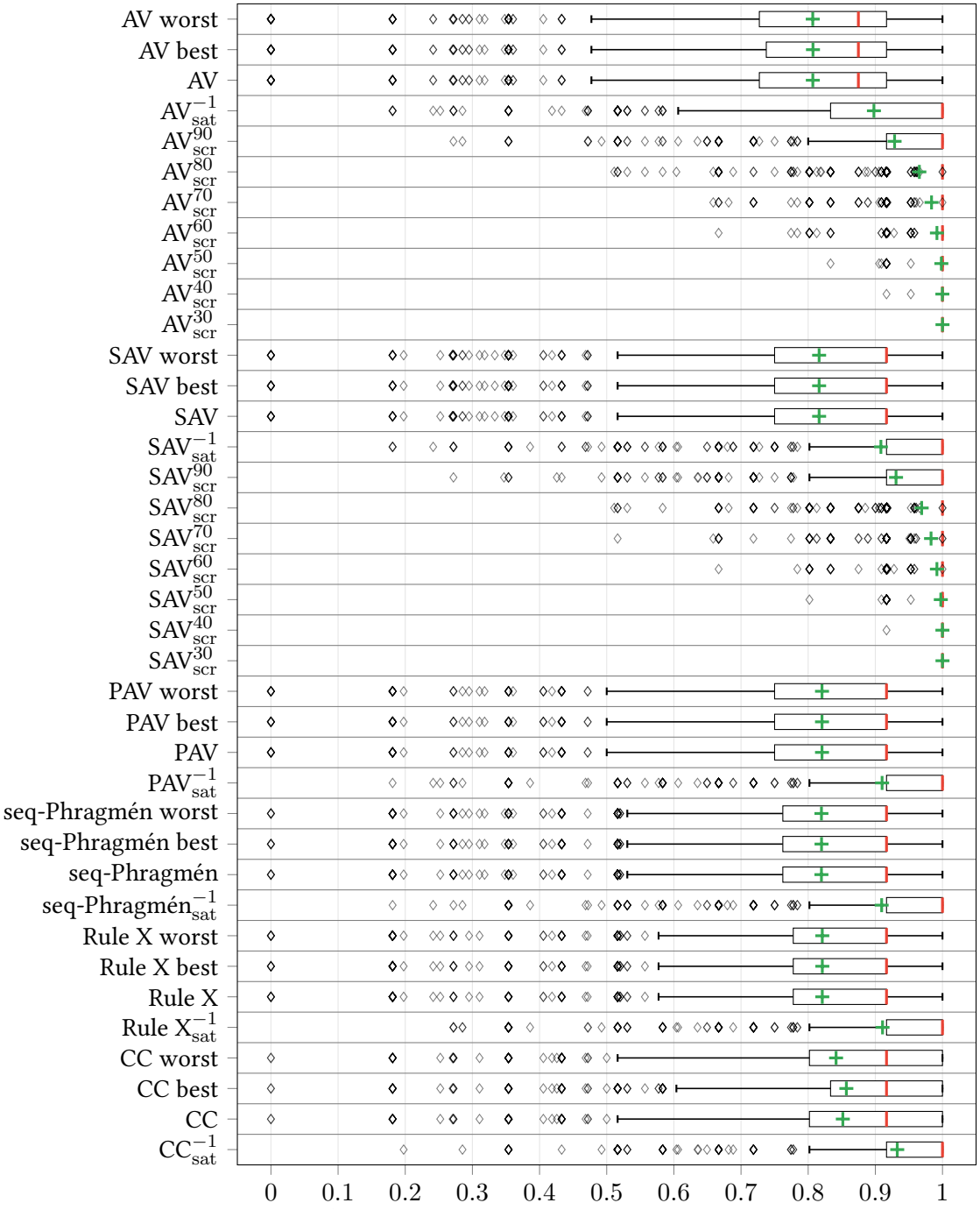


Figure 8: The proportion of the optimal diversity reached over all experimental data with $k = 8$ when using *Sh* for the different rules and approaches we consider. "$\mathcal{R}$ best" ("$\mathcal{R}$ worst") refers to the rule that chooses the committees with the highest (lowest) diversities among the winning committees of $\mathcal{R}$. If only $\mathcal{R}$ is written, the diversity of the winning committee that *abcvoting* returns for $\mathcal{R}$ is considered. The red line indicates the median, the green cross the mean.

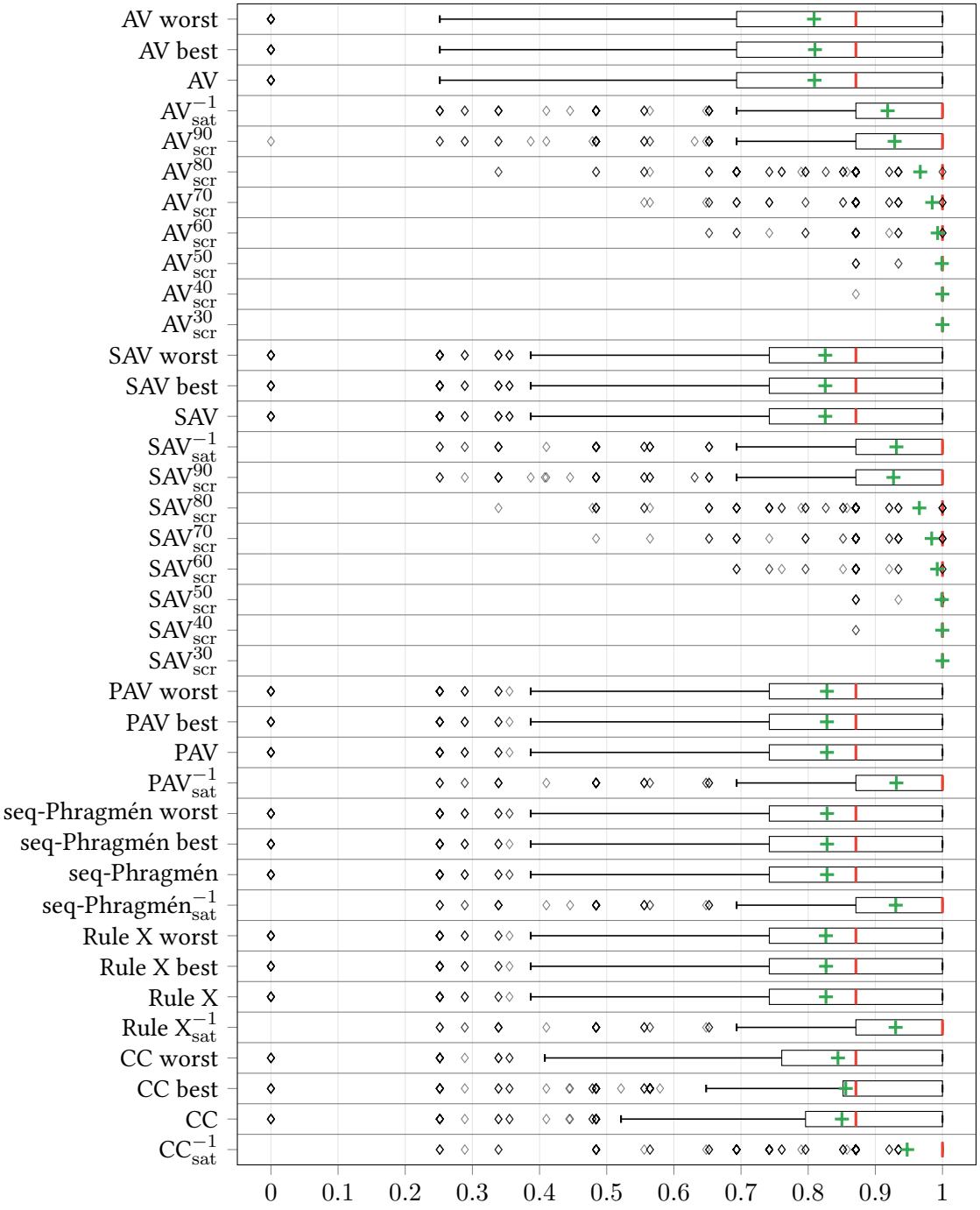


Figure 9: The proportion of the optimal diversity reached over all experimental data with $k = 6$ when using *Sh* for the different rules and approaches we consider. "$\mathcal{R}$ best" ("$\mathcal{R}$ worst") refers to the rule that chooses the committees with the highest (lowest) diversities among the winning committees of $\mathcal{R}$. If only $\mathcal{R}$ is written, the diversity of the winning committee that *abcvoting* returns for $\mathcal{R}$ is considered. The red line indicates the median, the green cross the mean.

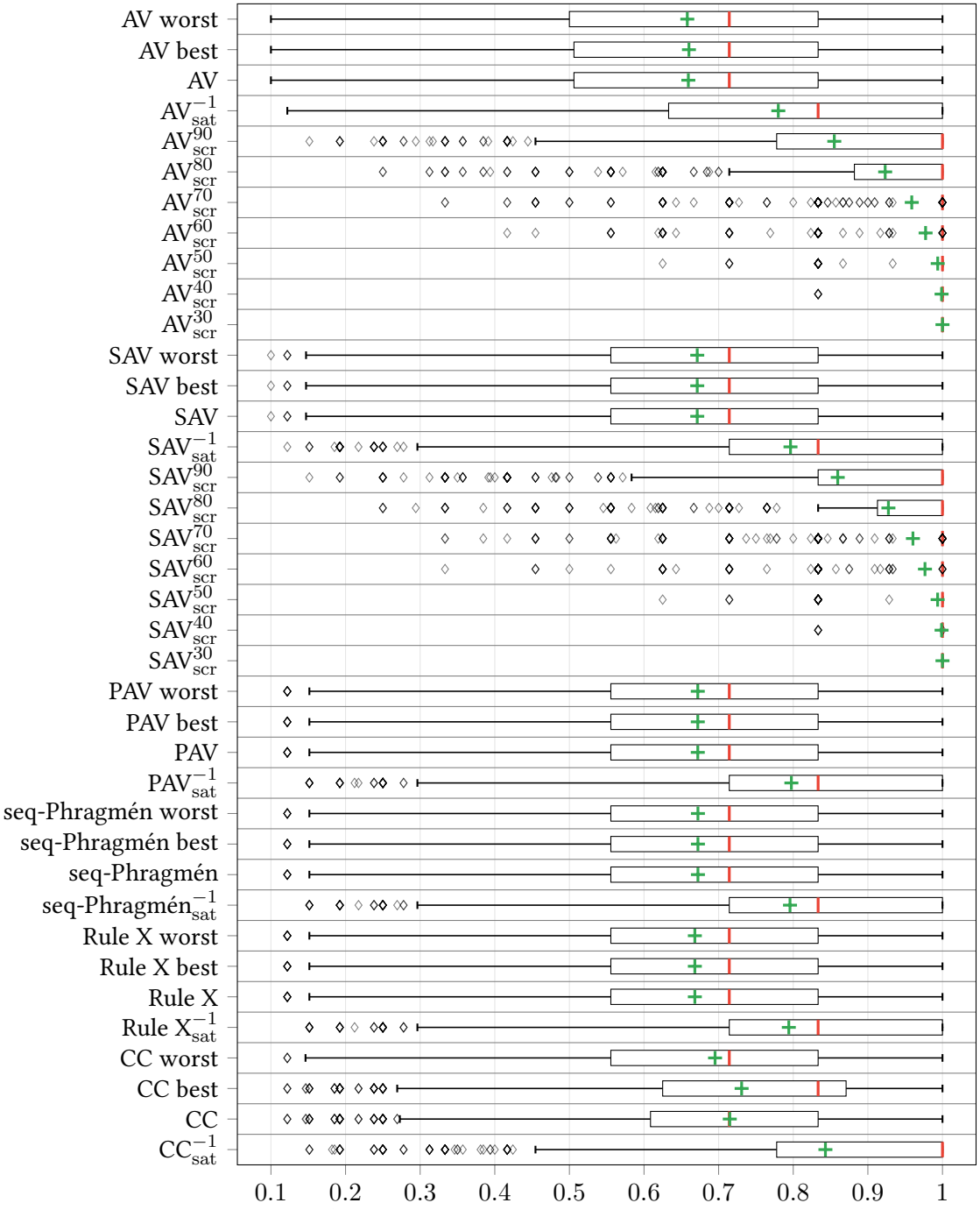


Figure 10: The proportion of the optimal diversity reached over all experimental data with $k = 10$ when using Si for the different rules and approaches we consider. "$\mathcal{R}$ best" ("$\mathcal{R}$ worst") refers to the rule that chooses the committees with the highest (lowest) diversities among the winning committees of $\mathcal{R}$. If only $\mathcal{R}$ is written, the diversity of the winning committee that *abcvoting* returns for $\mathcal{R}$ is considered. The red line indicates the median, the green cross the mean.

Figure 11: The proportion of the optimal diversity reached over all experimental data with $k = 8$ when using Si for the different rules and approaches we consider. "$\mathcal{R}$ best" ("$\mathcal{R}$ worst") refers to the rule that chooses the committees with the highest (lowest) diversities among the winning committees of $\mathcal{R}$. If only $\mathcal{R}$ is written, the diversity of the winning committee that *abcvoting* returns for $\mathcal{R}$ is considered. The red line indicates the median, the green cross the mean.

Figure 12: The proportion of the optimal diversity reached over all experimental data with $k = 6$ when using Si for the different rules and approaches we consider. "$\mathcal{R}$ best" ("$\mathcal{R}$ worst") refers to the rule that chooses the committees with the highest (lowest) diversities among the winning committees of $\mathcal{R}$. If only $\mathcal{R}$ is written, the diversity of the winning committee that *abcvoting* returns for $\mathcal{R}$ is considered. The red line indicates the median, the green cross the mean.

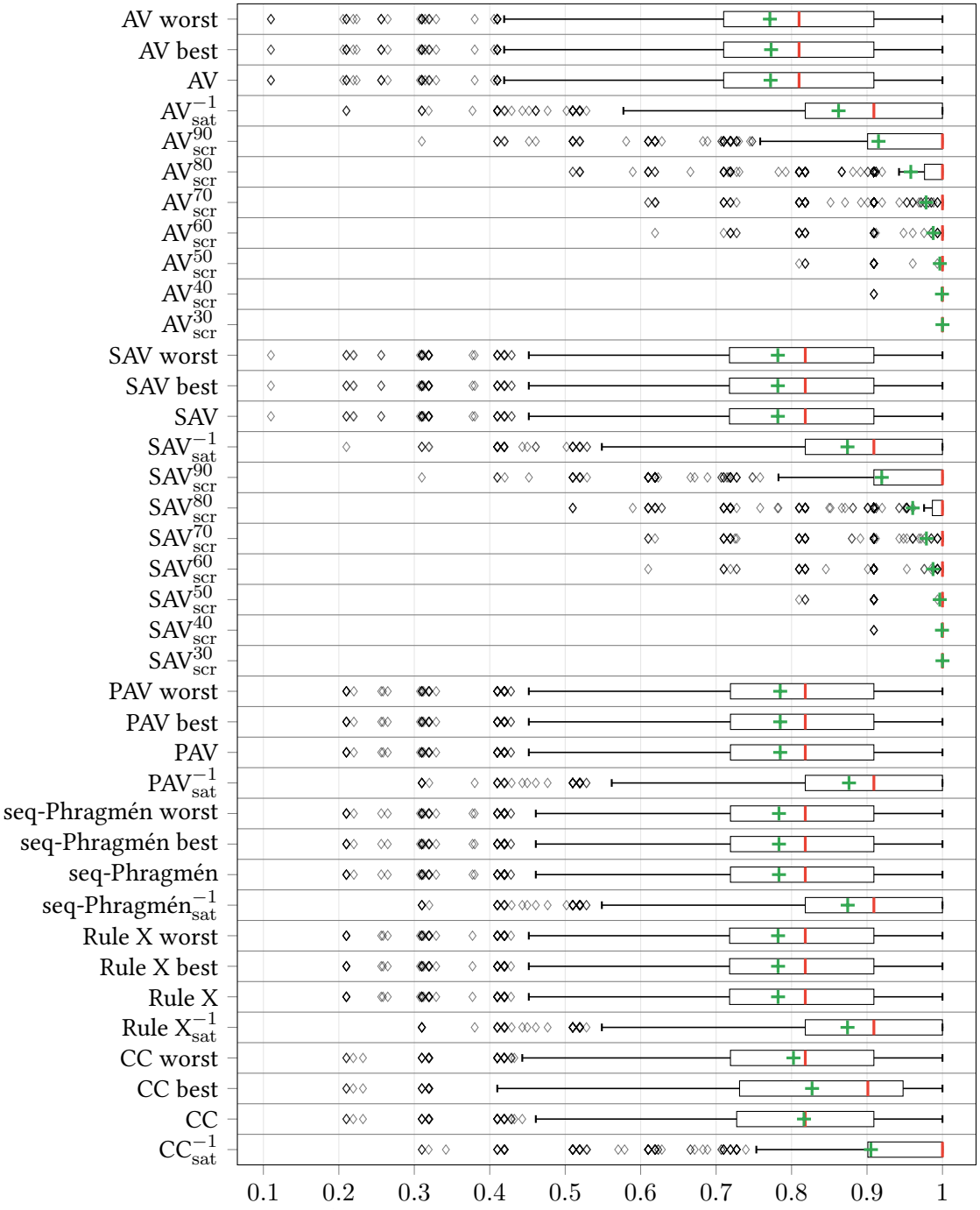


Figure 13: The proportion of the optimal diversity reached over all experimental data with $k = 10$ when using LC for the different rules and approaches we consider. "$\mathcal{R}$ best" ("$\mathcal{R}$ worst") refers to the rule that chooses the committees with the highest (lowest) diversities among the winning committees of $\mathcal{R}$. If only $\mathcal{R}$ is written, the diversity of the winning committee that *abcvoting* returns for $\mathcal{R}$ is considered. The red line indicates the median, the green cross the mean.

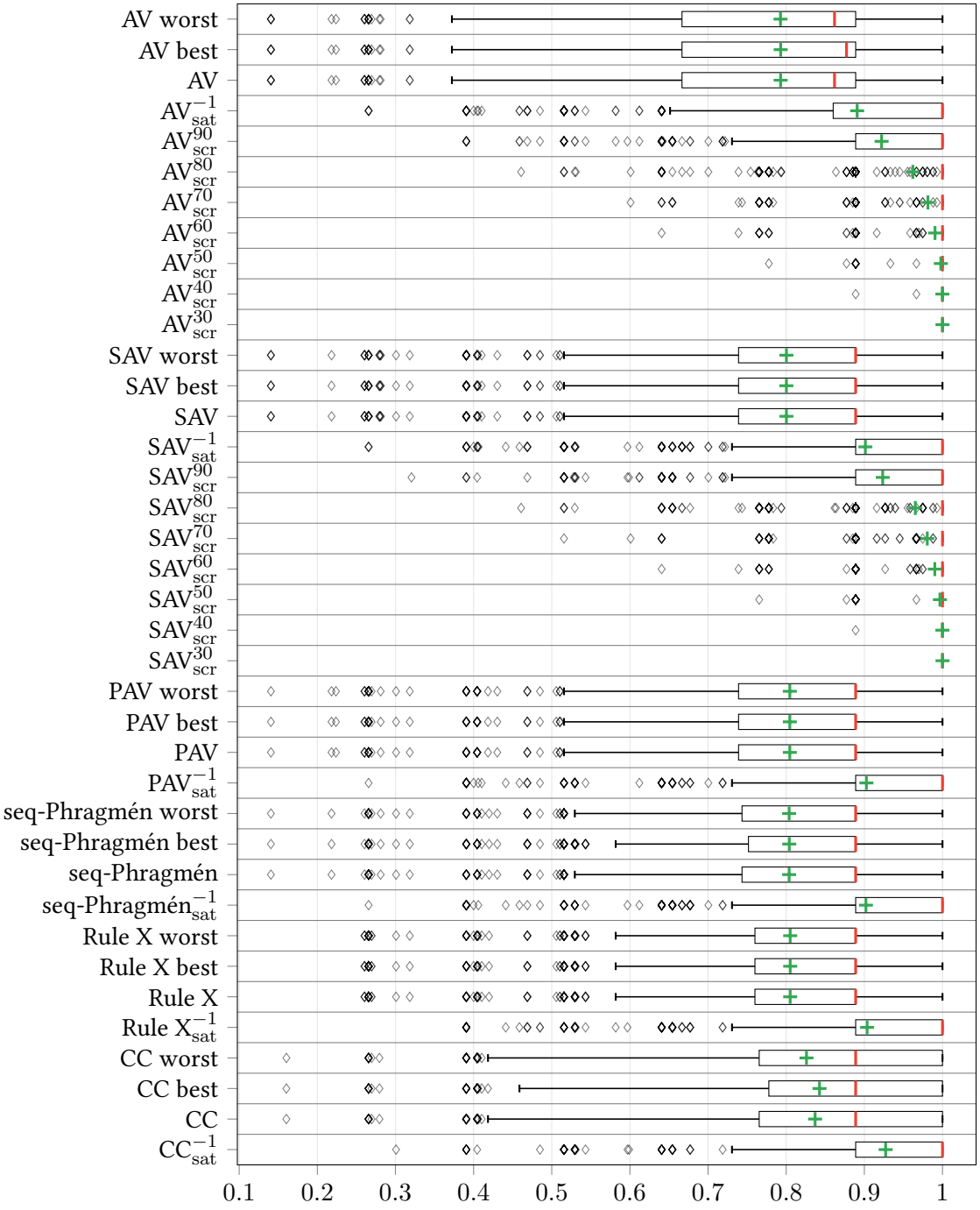


Figure 14: The proportion of the optimal diversity reached over all experimental data with $k = 8$ when using LC for the different rules and approaches we consider. "$\mathcal{R}$ best" ("$\mathcal{R}$ worst") refers to the rule that chooses the committees with the highest (lowest) diversities among the winning committees of $\mathcal{R}$. If only $\mathcal{R}$ is written, the diversity of the winning committee that *abcvoting* returns for $\mathcal{R}$ is considered. The red line indicates the median, the green cross the mean.

Figure 15: The proportion of the optimal diversity reached over all experimental data with $k = 6$ when using LC for the different rules and approaches we consider. "$\mathcal{R}$ best" ("$\mathcal{R}$ worst") refers to the rule that chooses the committees with the highest (lowest) diversities among the winning committees of $\mathcal{R}$. If only $\mathcal{R}$ is written, the diversity of the winning committee that *abcvoting* returns for $\mathcal{R}$ is considered. The red line indicates the median, the green cross the mean.