Diverse Committees with Incomplete or Inaccurate Approval Ballots

AAAI Track

Feline Lindeboom
Rijksuniversiteit Groningen
Groningen, The Netherlands
felinelindeboom@proton.me

Martijn Brehm
Universiteit van Amsterdam
Amsterdam, The Netherlands
m.a.brehm@uva.nl

Davide Grossi
Rijksuniversiteit Groningen
Universiteit van Amsterdam
Groningen, The Netherlands
d.grossi@rug.nl

Pradeep K. Murukannaiah
Technische Universiteit Delft
Delft, The Netherlands
p.k.murukannaiah@tudelft.nl

ABSTRACT

We study diversity in approval-based committee elections with incomplete or inaccurate information. We define diversity according to the Maximum Coverage problem, which is known to be NP-complete, with a best attainable polynomial time approximation ratio of $1 - 1/e$. In the incomplete information setting, voters vote only on a small portion of the candidates, and we prove that getting arbitrarily close to the optimal approximation ratio w.h.p. requires $\Omega(m^2)$ non-adaptive queries, where m is the number of candidates. This motivates studying adaptive querying algorithms, that can adapt their querying strategy to information obtained from previous query outcomes. In that setting, we lower this bound to only $\Omega(m)$ queries. We propose a greedy algorithm to match this lower bound up to log-factors. We prove the same $\tilde{\Theta}(m)$ bound for the generalized problem of Maximum Coverage over a matroid constraint, using a local search algorithm. Specifying a matroid of valid committees lets us implement extra structural requirements on the committee, like quota. In the inaccurate information setting, voters' responses are corrupted with a small probability. We prove $\tilde{\Theta}(nm)$ queries are required to attain a $(1-1/e)$-approximation with high probability, where n is the number of voters. While the proven bounds show that all our algorithms are viable asymptotically, they also show that some of them would still require large numbers of queries in instances of practical relevance. Using real data from Polis as well as synthetic data, we observe that our algorithms perform well also on smaller instances, both with incomplete and inaccurate information.

KEYWORDS

Computational Social Choice; Approval-Based Committee Elections; Chamberlin-Courant; Incomplete and Inaccurate Information

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/UDNL2582

ACM Reference Format:
Feline Lindeboom, Martijn Brehm, Davide Grossi, and Pradeep K. Murukannaiah. 2026. Diverse Committees with Incomplete or Inaccurate Approval Ballots: AAAI Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/UDNL2582

1 INTRODUCTION

Consensus has grown among political scientists, democracy practitioners, and decision makers alike, that effective involvement of citizens in policy decisions should be regarded as a high priority for democratic institutions at all levels, from local, to national, to regional [21, 25]. Currently, it is on digital democracy platforms especially that research efforts are concentrating [5, 14], in order to provide citizens with effective civic participation tools [26].

A wealth of these types of digital democracy tools have been developed and deployed around the world in the last decade: LiquidFeedback [2], Consul, YourPriorities, Decidim, Polis [35], to mention a few.[1] In different forms, all these applications allow users to provide free-text input to public deliberations and enable them to express their own opinions on the input of others. Supporting such processes involves the use of algorithms for information-processing problems. One such problem concerns how to effectively summarize the current state of the deliberation: *how does 'the group' currently think about the issues being deliberated upon?*

One approach to this problem, which is followed for instance in Polis, consists of selecting sets of statements proposed by users, based on how much support the statements have elicited from other users. As Halpern, Kehne, Procaccia, Tucker-Foltz and Wüthrich [15] already noticed, this approach can be conceptualized as an approval-based committee election problem [10, 20], as studied in computational social choice [4]: users 'vote' for other users' statements by expressing whether they approve/support them. Approaching the problem from this point of view has two main advantages. First, it can guarantee summaries that are representative of the discussion in a rigorous sense. Second, properly designed algorithms can enhance transparency of the decision-making process, which can in turn increase participants' trust in this same

[1] See www.liquidfeedback.org; www.consulproject.org; www.citizens.is; www.decidim.org; www.pol.is.

process [27]. Vice-versa a non-transparent process may generate distrust, while individuals with more political distrust tend to show less interest in politics and lower rates of civic participation [9].

Unlike in standard approval based committee elections, information about who supports which statements is now sparse, as no user can possibly express whether they agree or disagree with every single statement contributed by their peers. For the same reason, voters can become inaccurate in representing their true beliefs, when having to answer many questions. Such a deliberation summarization problem can thus be thought of as an approval-based committee election problem in which ballots are incomplete or inaccurate, or both. From this perspective, Halpern et al. [15] focused on the problem of designing selection algorithms yielding summarizations that meet specific forms of proportional representation proposed in the computational social choice literature [1], while querying users' opinions as efficiently as possible.

Our work builds on the approach put forth by Halpern et al. [15], but focuses on the property of diversity instead. Diversity has received less attention than proportional representation in the committee election literature but, we argue, is an important objective for deliberation summarization, as already mentioned in Lackner and Skowron [20]. *First*, diverse summaries can guarantee higher inclusivity, an attribute one could desire in and of itself. *Second*, an inclusive summary can show more participants that their input is taken into account. This increases people's trust in the system at hand [26], which motivates to participate also in future instances. *Third*, a diverse summary gives a broad, informative view of 'the group's' opinion. This is especially valuable in deliberation summarization contexts, where the number of candidates may be very large. *Fourth*, cognitive diversity is argued, e.g., by Landemore [21], to also be of epistemic value from a crowd-wisdom perspective. This is relevant because the selected statements typically serve as a basis for further discussions.

We use the established formalization of diversity in committee elections based on the Chamberlin-Courant score [7] and the Maximum Coverage problem [28, 34]. Our focus thus lies on the computational problem of constructing summarizations that maximize the coverage of voters, while querying users' opinions as efficiently as possible.

2 RELATED WORK AND CONTRIBUTION

Diverse committees and the Maximum Coverage problem. In committee election theory, the diversity of a committee is commonly quantified by the *Chamberlin-Courant* score [7]. In *approval-based* committee elections, a voter's *representative* in an elected committee W then is any one candidate in W they approve of, and the *approval-based Chamberlin-Courant* (going forward: Chamberlin-Courant) score of W equals the number of voters that have a representative in W. Portraying candidates by their set of approving voters, we see this problem amounts to selecting a set of candidates that maximizes the number of 'covered' voters: the approval-based Chamberlin-Courant problem is equivalent to the unit-weight Maximum Coverage problem.

An instance of the unit-weight Maximum Coverage problem, Max Cover going forward, consists of a set V of n elements v_i, a family $C \subseteq \mathcal{P}(V)$ of m subsets c_j, and a natural number k. The goal is to select k subsets so that $|\bigcup_{j=1}^{k} c_j|$ is maximized. Hochbaum and Pathria [16] proved, by reduction from Set Cover, that the Max Cover decision problem is NP-complete, and that a greedy algorithm is $(1-1/\mathrm{e})$-approximate. Shortly after, Feige [11], using a reduction from approximating Max 3SAT-5, proved that for any $\epsilon > 0$, Max Cover cannot be approximated in polynomial time within a ratio of $1-1/\mathrm{e}+\epsilon$, unless P = NP; the simple greedy algorithm is thus optimal. Many years later, Cohen-Addad, Gupta, Kumar, Lee and Li [8] showed that under the Gap Exponential Time Hypothesis[2], for any $\epsilon > 0$, there exists no FPT-approximation algorithm for parameter k that approximates Max Cover within factor $1-1/\mathrm{e}+\epsilon$. Manurangsi [23] improved the running time lower bound from $f(k) \cdot (m+n)^{k^{\mathrm{poly}(1/\epsilon)}}$ for any function f and $\epsilon > 0$, to $f(k) \cdot (m+n)^{o(k)}$, which is tight. Essentially, this shows a brute-force approach going over all size-k subsets is the best we can do if we insist on obtaining an approximation ratio better than $1-1/\mathrm{e}$.

Peters [30] proved that maximizing the Chamberlin-Courant score can be efficiently done on the Candidate Interval (CI) domain, an approval-based version of the Single-Peaked domain. Peters and Lackner [31] extended this result to circular preference domains and Sornat, Williams and Xu [36] extended it to nearly CI domains. Skowron and Faliszewski [34] made boundedness assumptions on the number of sets an element appears in. When any element appears in *at most* p sets, optimizing over the largest $\lceil \frac{2pk}{1-\beta} + k \rceil$ sets yields a β-approximate solution in FPT time (in k). When any element appears in *at least* p sets, the greedy algorithm achieves an improved approximation ratio of $1 - \mathrm{e}^{-\max\{\frac{pk}{m}, 1\}}$. Finally, a generalization of the problem was studied by Filmus and Ward [12], who presented a local search algorithm that is $(1-1/\mathrm{e})$-approximate for Max Cover over a matroid constraint. The greedy algorithm attains only ratio $1/2$ in this generalized setting.

Incomplete or inaccurate information. Single-winner voting with incomplete information received some attention over the years in the computational social choice literature [3]. More recent work has addressed the issue of achieving proportional representation in committee elections under incomplete information [13, 15, 17], or of maximizing arbitrary scoring rules for ranking candidates, subject to the satisfaction of the Justified Representation criterion [32]. Among these contributions, the work of Halpern et al. [15] is closest to what is our focus in this paper. The authors study proportionality in approval-based committee elections under incomplete information. The paper presents a version of local search Proportional Approval Voting that queries voters and, using $O(mk^6 \log k \log m)$ queries, finds a solution that satisfies Extended Justified Representation [1] and Optimal Average Satisfaction (based on the notion of average satisfaction [38]) with high probability. The result extends to an α-approximate version of both axioms, in which case the query complexity decreases by a factor k^3. The same study presents a lower bound of $\Omega(m^{11})$ for the query complexity of algorithms that cannot adapt their querying strategy to information obtained from previous queries (non-adaptive algorithms).

[2]For $k \geq 3$, define $s_k := \inf\{\delta : \exists$ an algorithm that solves k-SAT in $2^{\delta n}$ time$\}$. ETH states that for $k \geq 3$, $s_k > 0$; k-SAT cannot be solved in subexponential time, for $k \geq 3$. ETH was first formulated by Impagliazzo and Paturi [18] and would imply that P $\neq$ NP.

Table 1: Overview of query complexity bounds in m and n.

Incomplete			Inaccurate
Adaptive		Non-adaptive	$\tilde{\Theta}(nm)$
Matroid	No matroid	$\Omega(m^2)$	
$\tilde{\Theta}(m)$	$\tilde{\Theta}(m)$		

Our contribution. We study diversity in approval-based committee elections when information elicited from voters is incomplete or inaccurate. We measure the diversity of a solution with its Max Cover, or equivalently, Chamberlin-Courant, score. For a large part, our work combines the two lines of research outlined above, and is based in particular on the work of Filmus and Ward [12] and Halpern et al. [15]. We make three main contributions.

First, in a setting with incomplete information, we prove that getting arbitrarily close to the optimal approximation ratio with high probability (w.h.p.) requires $\Omega(m^2)$ non-adaptive queries (this result is presented only in the full version of the paper[3]). This motivates studying *adaptive* querying algorithms, that can adapt their querying strategy to information obtained from previous query outcomes. In that setting, we lower this bound to only $\Omega(m)$ queries. We adapt the greedy algorithm to match this lower bound up to log-factors (Theorem 2). We prove the same $\tilde{\Theta}(m)$[4] bound for the generalized problem of Max Cover over a matroid constraint (Theorem 5). Specifying a matroid of valid committees lets us implement external diversity requirements, like upper and lower quota on groups of candidates. *Second*, in the inaccurate information setting, we prove that recovering the optimal approximation ratio w.h.p. requires $\tilde{\Theta}(nm)$ queries (Theorem 7). Despite these positive asymptotic bounds, summarized in Table 1, our results do involve sizeable constant overheads for some of our algorithms, which make viability on real-world instances questionable. So, *third*, using real data from Polis and synthetic data that we produced by adapting methods from Szufa, Faliszewski, Janeczko, Lackner, Slinko, Sornat and Talmon [37], we empirically show that our algorithms perform considerably better than our worst-case analysis suggests. Importantly, this appears to hold even in situations in which votes are both incomplete *and* inaccurate. All proofs are provided in the full version of the paper.

3 PRELIMINARIES

An instance of an approval-based committee election problem consists of a set V of n voters, a set C of m candidates, and a natural number k. The goal is to elect a committee $W \subseteq C$ of size $k \leq m$, based on the opinions of the voters in V. An instance also contains approval information: every voter approves of a subset of the candidates. Usually, this approval information is known upfront and expressed in the form of an *approval set* $A(i)$ for each voter $v_i \in V$. The sequence of sets $A = (A(1), \ldots, A(n))$ is called the *approval profile*. We call this setting the *perfect information setting*. This paper studies incomplete and inaccurate information settings; the respective models are detailed in the corresponding section. To measure the diversity of a committee, we use the Chamberlin-Courant score.

Definition 1 (Chamberlin and Courant [7]). *On any approval-based committee election instance (V, C, A, k), the* Chamberlin-Courant (CC) *score of a committee $W \subseteq C$, $|W| = k$ is*

$$CC(W) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{A(i) \cap W \neq \emptyset\}}(i). \tag{1}$$

In the Chamberlin-Courant decision problem, the input is an instance (V, C, A, k) and an integer x, and the question is whether it is possible to achieve score $\frac{x}{n}$ using k candidates. In the Chamberlin-Courant problem, the input is an instance (V, C, A, k) and the task is to maximize the score using k candidates.

As previously mentioned, the approval-based Chamberlin-Courant problem is equivalent to the unit-weight Maximum Coverage problem. For the rest of the paper, we use the terminology of the Chamberlin-Courant optimization problem, with voters, candidates and committees (as opposed to elements, sets and covers).

4 INCOMPLETE INFORMATION

In the *incomplete information setting*, the approval profile $A = (A(1), \ldots, A(n))$ is replaced by a function $A(i, j) := \mathbb{1}_{\{c_j \in A(i)\}}(i, j)$ which equals 1 in case voter v_i approves candidate c_j and 0 otherwise. This is a generalization of the original model, as querying all voters about all candidates recovers the complete approval profile. As such, by making nm queries to A we can find a $(1 - 1/\mathrm{e})$-approximate solution, e.g. with a greedy approach. In the full version of the paper, we show that any non-adaptive querying algorithm must make $\Omega(m^2)$ queries to A to get arbitrarily close to this optimal approximation ratio w.h.p., where non-adaptive means that the sequence of queries made must be specified beforehand. In Polis data we observe that $m \in \Theta(n)$, which would imply $\Omega(m^2) = \Omega(nm)$, making no improvement over the complete information setting. We can do better by studying adaptive querying algorithms, that can adapt their querying strategy to information obtained from previous query outcomes. For such adaptive algorithms we show that obtaining an approximation ratio arbitrarily close to this optimal ratio w.h.p. requires only $\tilde{\Theta}(m)$ queries to A.

To see that we need at least $\Omega(m)$ queries, imagine an instance where only one candidate has any approvals. Any algorithm must select this candidate to attain a positive approximation ratio. In the worst-case, finding this candidate w.h.p. requires sampling voters for each of the m candidates. In the remainder of the section we study algorithms that witness this lower bound $\Omega(m)$ up to log-factors. We first adapt a greedy algorithm for CC and then do the same for a local search algorithm for the generalized problem of optimizing over a matroid constraint. Both algorithms achieve an approximation ratio arbitrarily close to the optimal ratio w.h.p. using $\tilde{O}(m)$ queries, matching the lower bound (up to log-factors).

4.1 Upper bound for the unconstrained setting

The standard greedy algorithm. We prepare the ground by considering the greedy algorithm for the perfect information setting. See Algorithm 1, where we write $\Delta(W, c)$ for the increase in CC-score

[3] https://arxiv.org/abs/2506.10843

[4] The notation $\tilde{\Theta}(\cdot)$ and similarly for Ω, O, suppresses terms of the form $\log^{O(1)}(n)$ where n is the growing parameter.

obtained by adding candidate c to committee W. Algorithm 1 elects, in each iteration, the candidate that yields the largest immediate increase in CC-score, and achieves the optimal approximation ratio of $1 - 1/e$ [16]. We give a novel proof of its approximation ratio in the full version of this paper.

The greedy query algorithm. We adapt Algorithm 1 to the incomplete information setting. Our strategy is to repeatedly sample sets of voters (of size $\ell \leq n$) uniformly at random and query them about a subset of candidates of size $t \leq m$, in order to estimate $\Delta(W, c)$ for all $c \notin W$. The algorithm then behaves like the standard greedy algorithm: iteratively selecting the candidate c maximizing this estimate of $\Delta(W, c)$. By carefully bounding the deviation from the true value $\Delta(W, c)$, we can guarantee that our algorithm attains a ratio arbitrarily close to the optimal approximation ratio w.h.p.

To write this more formally, note that if we ask *all* voters for their votes on some query set $Q \subseteq C$ (of size t), the responses allow us to compute, for any set $S \subseteq Q$:

$$p_S := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{A(i) \cap S \neq \emptyset\}}(i). \tag{2}$$

The value p_S equals the Chamberlin-Courant score of S, and in this case $\Delta(W, c) = p_{W \cup \{c\}} - p_W$, so we can compute $\Delta(W, c)$ with complete information on a query set containing $W \cup \{c\}$. Sampling ℓ voters uniformly at random and querying them about the set $Q \subseteq C$, we can compute, for any subset $S \subseteq Q$,

$$\hat{p}_S := \frac{1}{\ell} \sum_{i=1}^{\ell} \mathbb{1}_{\{A(i) \cap S \neq \emptyset\}}(i), \tag{3}$$

that estimates p_S. We write $\hat{\Delta}(W, c) := \hat{p}_{W \cup \{c\}} - \hat{p}_W$ for an estimate of $\Delta(W, c)$ based on $\hat{p}$. Using this approach, we obtain Algorithm 2.

In step 1 of the for-loop, to be able to calculate the values $\Delta(W, c')$, we add the committee to each query set and, other than that, distribute the candidates over the query sets until they reach size t, without further restrictions.

For Theorem 2, we assume that $\sum_{v_i \in V} \mathbb{1}_{\{|A(i)| \geq 1\}}(i) \geq k$, that is: there exist at least k voters with a non-empty approval set.[5] Theorem 2 gives an upper bound on the number of queries required to run Algorithm 2 on such instances.

[5]We use this assumption to prove the approximation guarantee of the greedy query algorithm, see also the full version of the paper. The assumption is also motivated practically: instances where $\sum_{v_i \in V} \mathbb{1}_{\{|A(i)| \geq 1\}}(i) < k$ are uninteresting in the context of deliberation summarization—where the challenge is to aggregate *many* different opinions into a small summary—and do not occur in practice.

ALGORITHM 1: GREEDY

Input: Numbers $n, m, k \in \mathbb{N}$ with $k \leq m$, set V of n voters, set C of m candidates.
Output: Committee $W \subseteq C$ of size k.
Let $W = \{\}$ be an empty set;
for $i = 1, \ldots, k$, **do**
  Add $c' \in \operatorname{argmax}_{c \notin W} \Delta(W, c)$ to W.
end

ALGORITHM 2: GREEDY-INCOMPLETE

Input: Numbers $n, m, k \in \mathbb{N}$ with $k \leq m$, set V of n voters, set C of m candidates, query size t with $m \geq t > k$, and $\gamma \in (0, 1)$, $\delta > 0$.
Output: Committee $W \subseteq C$ of size k.
Set $\epsilon = \frac{(1-\gamma)e}{\gamma(e-1)}$ and $\ell = \lceil \frac{2}{\epsilon^2} (\log(\frac{2mk}{\delta})) \rceil$;
Let $W = \{\}$ be an empty set;
for $i = 1, \ldots, k$, **do**
  (1) Construct the smallest set of query sets $\mathcal{Q} = \{Q_i\}_i$ with $Q_i \subseteq C$ and $|Q_i| = t$ for all i, and such that $W \subseteq \bigcap \mathcal{Q}$ and $\bigcup \mathcal{Q} = C$, and present each query set to ℓ voters sampled uniformly at random.
  (2) For all $c \notin W$, determine $\hat{\Delta}(W, c)$ as an estimate of $\Delta(W, c)$ using ℓ voters and query set Q containing $W \cup c$.
  (3) Add $c' \in \operatorname{argmax}_{c \notin W} \hat{\Delta}(W, c)$ to W.
end

THEOREM 2. *Let $\sum_{v_i \in V} \mathbb{1}_{\{|A(i)| \geq 1\}}(i) \geq k$, $\delta > 0$, $\gamma \in (0, 1)$, and $k < t \leq m$. Then, w.p. at least $1 - \delta$, Algorithm 2 is $(1 - 1/e)\gamma$-approximate for CC with query complexity*

$$O\left(\left(\frac{\gamma}{1-\gamma}\right)^2 km \log\left(\frac{km}{\delta}\right)\right) \in O_{\delta,\gamma,k}(m \log m).$$

A few observations are in order. First note that, asymptotically, the query complexity provided by Theorem 2 is much smaller than querying the entire profile of size nm. The constant overhead, which increases with k and γ and decreases with δ, is quite small. Second, note that a voter may be queried more than once during the run of the algorithm, because we sample with replacement. However, since the query complexity is independent of n, sampling with replacement approaches sampling without replacement as n grows. Third, the asymptotic complexity does not depend on t. Increasing t means fewer voters get larger query sets, which decreases the total number of queries only slightly (by a small constant factor). In practice, the value of t should be chosen large enough so that the query sets leave room for sufficiently many unelected candidates, but not too large, since we assume that voters are not able to express their opinion on all m candidates.

To give further context to Theorem 2: in real-world instances of online deliberations such as those ran through Polis, we observe that typically $m \in \Theta(n)$ and $k \ll m$. So, suppose that $m = 1000$, $\gamma = 0.85$, $\delta = 0.05$, $k = 8$ and $t = 20$, then Theorem 2 requires 432 920 queries. Starting with $n > 432$ voters this is smaller than nm.

4.2 Upper bound for a setting with matroid constraints

Matroid constraints for committee elections. We can imagine situations where we want to place restrictions on the set of possible committees to be elected. Perhaps we want to include extra diversity requirements, such as 'the committee must contain at least/at most x candidates of category Y'. In practice, x would be a natural number and Y could be a demographic group. In online deliberation settings, these attributes would extend to users. This provides the possibility to add an extra 'dimension' of diversity: where before, the concept of diversity was based solely on the approval profile, we now consider *external attributes* of the candidates. For such purposes, we make a selection of which committees are 'allowed'

to be elected, and collect them in a set $\mathcal{I}$. It is not obvious that any algorithm adapts well to this restriction. Fortunately, when $\mathcal{I}$ defines a matroid, we can mend this problem [12], and, as shown by Masařík, Pierczyński and Skowron [24], matroids can indeed also be used to implement both upper and lower quota on any number of disjoint categories. See the full version of the paper for the construction.

Definition 3 (matroid). *A matroid $\mathcal{M}$ is a pair $(C, \mathcal{I})$ where C is a finite universe and $\mathcal{I}$ is a collection of subsets of C (called independent sets) satisfying the following properties:*

(1) $\emptyset \in \mathcal{I}$,
(2) *if $A \in \mathcal{I}$ and $B \subset A$ then $B \in \mathcal{I}$, and*
(3) *for all $A, B \in \mathcal{I}$ with $|A| > |B|$, there exists $x \in A \setminus B$ such that $B \cup \{x\} \in \mathcal{I}$.*

Property 3) guarantees that all maximal independent sets have the same cardinality. We call such sets *bases*, and their common cardinality is called the *rank* of the matroid.

Example 4 (uniform matroid). *Consider $\mathcal{I} = \{W \subseteq C : |W| \leq k\}$. Then $\mathcal{M} = (C, \mathcal{I})$ is the uniform matroid of rank k: 1) $|\emptyset| = 0 \leq k$, 2) for any set $W \in \mathcal{I}$ with $|W| = j \leq k$, for any subset $W' \subset W$, $|W'| < |W| \leq k$, and 3) for any two feasible sets W, $W' \in \mathcal{I}$ with $|W'| < |W|$, we will have, for any $c \in W \setminus W'$, that $|W' \cup \{c\}| \leq k$.*

Thus, taking for $\mathcal{M}$ the uniform matroid of rank k, we retrieve the original problem: the problem with matroid constraints is a generalization of the original problem.

The standard non-oblivious local search algorithm. We start with the perfect information setting. A classical (or oblivious) local search algorithm starts with an arbitrary solution and, in each iteration, swaps one or multiple of the elected elements for unelected elements in order to improve the objective function value. It thus *searches* in the *local* neighborhood of the solution. It stops when no local improvement is possible. At worst, this occurs after all exponentially many options have been exhausted. With the use of approximate local search, this running time problem can be resolved at a small cost in the approximation ratio [29]. In practice, this means we terminate the algorithm when we encounter an improvement smaller than some parameter $\beta > 0$[6].

Over a matroid constraint, the greedy algorithm attains only approximation ratio $\frac{1}{2} < 1 - 1/\mathrm{e}$. The approximation ratio of oblivious local search is $\frac{k-1}{2k-\ell-1}$ when ℓ sets are exchanged in each iteration [12]. For $k = 2$ and $\ell = 1$, this also equals $\frac{1}{2}$. Filmus and Ward [12] thus switch to a non-oblivious local search algorithm that uses an auxiliary objective function f for the iterative procedure. The function adds temporary weight to elements covered multiple times, the idea being that elements covered multiple times, remain covered after the next exchange. Such a function can thus prevent getting stuck in bad local optima. We denote by α_j the temporary weight associated to an element covered j times, and write $h_i(W)$ for the number of times that v_i is covered by W. The auxiliary objective function is then defined as

$$f(W) = \frac{1}{n} \sum_{i=1}^{n} \alpha_{h_i(W)}. \tag{4}$$

Note that, with $\alpha_0 = 0$ and $\alpha_j = 1$ for all $j > 0$, we retain the original oblivious objective function, but if we set $\alpha_j > \alpha_1$ for $j > 1$, we add additional weight to elements covered multiple times. We define

$$\alpha_0 = 0, \quad \alpha_1 = 1 - \frac{1}{\mathrm{e}}, \quad \alpha_{j+1} = (j+1)\alpha_j - j\alpha_{j-1} - \frac{1}{\mathrm{e}}.$$

This choice for the sequence $(\alpha_n)_{n \in \mathbb{N}_0}$ is optimal and, for any $\gamma \in (0, 1)$, with this objective function, the non-oblivious local search algorithm, with parameter β (decreasing in γ) is $(1 - 1/\mathrm{e} - \gamma)$-approximate and runs in polynomial time [12].

We adjust the non-oblivious local search algorithm of Filmus and Ward [12] to suit our setting better, see Algorithm 3, where for $c \in W$ and $c' \notin W$, we now write $\Delta(W, c', c) := f((W \cup \{c'\}) \setminus \{c\}) - f(W)$. We elaborate on the adaptations in the full version of the paper, but most importantly, the changes do not affect the approximation guarantees of the algorithm: for $\beta = C_1 \frac{\gamma}{k \log k}$, Algorithm 3 is $(1 - 1/\mathrm{e} - \gamma)$-approximate for any $\gamma \in (0, 1)$ and some universal constant $C_1 \leq \frac{\log i}{\alpha_i (1 - 1/\mathrm{e} - \gamma)}$ for any $i \leq k$ (see also [12, Corollary 6]).

ALGORITHM 3: local search-β

Input: Numbers $n, m, k \in \mathbb{N}$ with $k \leq m$, set V of n voters, set C of m candidates, $\beta > 0$, matroid $\mathcal{M}$.
Output: Committee $W \subseteq C$, of size k.
Choose $W \subseteq C$ such that $|W| = k$, and $c \in W$ and $c' \notin W$ so that $(W \cup \{c'\}) \setminus \{c\} \in \mathcal{I}$;
repeat
(1) $W = (W \cup \{c'\}) \setminus \{c\}$.
(2) Let $\mathcal{E}$ be the set of all valid exchanges for W according to $\mathcal{M}$.
(3) Pick $(c', c) \in \operatorname{argmax}_{(x,y) \in \mathcal{E}} \Delta(W, x, y)$.
until $\Delta(W, c', c) \leq \beta$;

The non-oblivious local search query algorithm. We turn to the incomplete information setting. Recall that $A(i, j)$ equals 1 when voter v_i approves of candidate c_j and 0 otherwise, and that we write $A(i)$ for the set of candidates approved by voter v_i. Given a query set $Q \subseteq C$, for any $S \subseteq Q$, we redefine

$$p_S := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{A(i) \cap Q = S\}}(i); \tag{5}$$

the probability that a uniformly chosen voter approves, of all the candidates in Q, exactly the set S. This is different from how we defined p_S for Algorithm 2: Algorithm 2 requires, for any voter v_i, knowing just whether they have a representative in set $S \subseteq Q$, whereas Algorithm 4 needs the number of representatives. With p_S for $S \subseteq Q$, we can compute $f(Q)$:

$$f(Q) = \frac{1}{n} \sum_{i=1}^{n} \alpha_{h_i(Q)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{S \subseteq Q} \mathbb{1}_{\{A(i) \cap Q = S\}}(i) \cdot \alpha_{|S|}$$

$$= \sum_{S \subseteq Q} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{A(i) \cap Q = S\}}(i) \cdot \alpha_{|S|} = \sum_{S \subseteq Q} p_S \cdot \alpha_{|S|}.$$

[6] Using a partial enumeration technique, we could even eliminate this small cost again, see Calinescu, Chekuri, Pál and Vondrák [6].

ALGORITHM 4: LS-β-INCOMPLETE

Input: Numbers $n, m, k \in \mathbb{N}$ with $k \leq m$, set V of n voters, set C of m candidates, $\beta > 0$, query size t, with $m \geq t > k$, constants $\delta > 0$ and $\xi \geq 1$, matroid $\mathcal{M}$.

Output: Committee $W \subseteq C$, of size k.

$\epsilon = \frac{\xi-1}{2\xi} \cdot \beta$, $\ell = \left\lceil \frac{(2-2/e)^2}{2\epsilon^2} \log\left(\frac{2\cdot(m-k)k\cdot\xi\alpha_k}{\delta\cdot\beta}\right)\right\rceil$;

Choose $W \subseteq C$ such that $|W| = k$, and $c \in W$ and $c' \notin W$ so that $(W \cup \{c'\}) \setminus \{c\} \in \mathcal{I}$;

repeat

(1) $W = (W \cup \{c'\}) \setminus \{c\}$;
(2) Let $\mathcal{E}$ be the set of all valid exchanges for W according to $\mathcal{M}$.
(3) Construct the smallest set of query sets $\mathcal{Q} = \{Q_i\}_i$ with $Q_i \subseteq C$ and $|Q_i| = t$ for all i, and such that $W \subseteq \bigcap \mathcal{Q}$ and $\bigcup \mathcal{Q} = C$, and present each query set to ℓ voters sampled uniformly at random.
(4) For all $c \in W$, $c' \notin W$, determine $\hat{\Delta}(W, c', c)$ as an estimate of $\Delta(W, c', c)$ using ℓ voters and query set Q containing $W \cup \{c'\}$.
(5) Pick $(c', c) \in \operatorname{argmax}_{(x,y)\in\mathcal{E}} \hat{\Delta}(W, x, y)$.

until $\hat{\Delta}(W, c', c) < \beta - \epsilon$;

We can compute $\Delta(W, c', c)$ with complete information on a query set containing $W \cup \{c'\}$. When information is incomplete, we estimate: sampling $\ell \leq n$ voters uniformly at random and presenting them with query set $Q \subseteq C$, we can compute, for every subset $S \subseteq Q$,

$$\hat{p}_S := \frac{1}{\ell} \sum_{i=1}^{\ell} \mathbb{1}_{\{A(i)\cap Q = S\}}(i), \tag{6}$$

that estimates p_S. We write $\hat{\Delta}(W, c', c) = \hat{f}((W \cup \{c'\}) \setminus \{c\}) - \hat{f}(W) = \sum_{S\subseteq(W\cup\{c'\})\setminus\{c\}} \hat{p}_S \cdot \alpha_{|S|} - \sum_{S\subseteq W} \hat{p}_S \cdot \alpha_{|S|}$ for an estimate of $\Delta(W, c', c)$ based on the values $\hat{p}$. Algorithm 4 is our local search query algorithm.

Theorem 5 assumes $|c_j| \geq 1$ for all j (all candidates have at least one approval) and $k \geq 3$[7]. It gives an upper bound on the number of queries required to run Algorithm 4 on such instances. As expected, this is higher than the bound stated in Theorem 2.

THEOREM 5. *Let $|c_j| \geq 1$ $\forall j$, $\delta > 0$, $\gamma \in (0, 1)$, $m \geq t > k \geq 3$. Fix $\beta = C_2 \frac{1-\gamma}{\gamma k \log k}$ for some constant C_2. Then, w.p. at least $1 - \delta$, Algorithm 4 is $(1 - 1/e)\gamma$-approximate for CC with query complexity $O\left(\left(\frac{\gamma k \log k}{1-\gamma}\right)^3 m \log\left(\frac{m\gamma k^2 \log k}{\delta(1-\gamma)}\right)\right) \in O_{\delta,\gamma,k}(m \log m)$.*

Like with GREEDY-INCOMPLETE, the query complexity does not depend on n, and is asymptotically much smaller than querying the entire ballot. Unlike before however, there is sizeable constant overhead increasing with k and γ, and decreasing with δ. Especially the dependence in k is much stronger than before. For the same example, $m = 1000$, $\gamma = 0.85$, $\delta = 0.05$, $k = 8$ and $t = 20$, Theorem 5 requires $9.52180 \cdot 10^{11}$ queries. Only for $n > 9.52180 \cdot 10^8$ this is smaller than nm.

[7] These conditions are sufficient to guarantee the claimed approximation ratio, and dropping them complicates the proofs unnecessarily, since both are very natural assumptions: candidates with no approvals can be removed since their election cannot influence the score, and committees of size $k = 1, 2$ are arguably uninteresting in the context of deliberation summarization. See also the full version of the paper.

5 INACCURATE INFORMATION

In the inaccurate information setting, voters do hand in complete ballots, but the voters are inaccurate in reporting their true approvals: we assume each query $A(i, j)$ is incorrect with a probability p. We model this by defining, for $p \in (0, \frac{1}{2})$, a *p-inaccurate query* as $A_p(i, j) := A(i, j) \oplus X$, where $X \sim$ Bernoulli(p)[8]. Samples of X are independent between voters, between candidates, and between consecutive draws of the same query. If $p = 0$, we would have accurate information, and with nm queries we would find a $(1 - 1/e)$-approximate solution (e.g. using the greedy algorithm). When answers may be corrupted, we can simply pose every question multiple times to compensate for the uncertainty. We do this to obtain an upper bound on the number of queries required to still acquire a $(1 - 1/e)$-approximate solution with high probability, adapting Algorithm 1 (see the full version of the paper). Moreover, we provide a matching lower bound (up to log-factors), using a result from multi-armed bandit theory, as done in Theorem 9 in [22]. Both results also hold for the problem of optimizing over a matroid constraint.

THEOREM 6. *Let $p \in (0, \frac{1}{2})$, $\delta > 0$, $n, m \in \mathbb{N}$. Then there exists an algorithm that is $(1 - 1/e)$-approximate for CC in the p-inaccurate query model w.p. at least $1 - \delta$ and with query complexity $O(nm \log(nm/\delta))$.*

THEOREM 7. *Let $p \in (0, \frac{1}{2})$, $\delta > 0$, $n, m \in \mathbb{N}$. Then any algorithm that is $(1 - 1/e)$-approximate for CC in the p-inaccurate query model w.p. at least $1 - \delta$ has expected query complexity $\Omega(nm \log(1/\delta))$.*

The two bounds match (up to log-factors). Observe that, unlike before, we assume access to the entire ballot, making the bound depend on n.

6 EXPERIMENTS

Motivation. We run our algorithms on real-life data and on synthetic data with realistic structure. We draw two main conclusions from these experiments.[9]

First, we establish empirically that the greedy and local search algorithms with complete information (Algorithm 1 and 3) consistently achieve a higher CC-score than two well known committee election algorithms, i.e. APPROVAL VOTING (selecting the k most popular candidates) and the LOCAL SEARCH PAV algorithm of Halpern et al. [15]. We take this to justify the study of our querying algorithms, as opposed to querying versions of these other algorithms, since the querying algorithms approximate the performance of the complete information algorithms.

Second, although we have proven that our querying algorithms can overcome inaccurate and incomplete voter responses with constant overhead, this constant overhead can still be very large. We show that in practice our algorithms beat the optimal approximation ratio, even when using much fewer queries than theoretically required, and even when we combine inaccurate and incomplete information (a setting that we did not study theoretically).

[8] This is essentially the classification noise model from PAC learning [19] applied to our setting.

[9] All code and data is available at https://github.com/martijnmartijnmartijnmartijn/Diverse-Committees-with-Incomplete-or-Inaccurate-Approval-Ballots.

Data. We test our algorithms on 18 open-use datasets of Polis deliberations.[10] As anticipated, participants typically vote on only a small portion of the statements, so we obtain a sparse comment-participant-matrix. Since these discussions have been completed already, we are not able to query voters anymore, so, for the sake of running our querying algorithms, we need to artificially complete the data. For details on this and other pre-processing steps we implemented, we refer to the full version of the paper.

As the 18 data sets available are too few to support robust observations, we generate synthetic data that are structurally similar to the Polis data. To do so, we turn to established methods for sampling approval-based elections, introduced by Szufa et al. [37]. Specifically, we employ the (q, ϕ)-resampling model to sample approval elections which, among all models discussed by Szufa et al. [37], provided the best fit with respect to the 18 data sets available. In this model, $q \in [0, 1]$ represents the fraction of approvals and $\phi \in [0, 1]$ represents the spread of approvals, so that $\phi = 0$ means all voters are identical and approve the exact same $\lfloor q \cdot m \rfloor$ candidates, while $\phi = 1$ means each candidate is approved with probability q independently so that the spread of approvals is maximal. After pre-processing the Polis data, we find $q = 0.0891$ and $\phi = 0.693$. The full version of the paper contains an explanation of how to arrive at these values. We sampled 100 elections according to the $(0.0891, 0.693)$-resampling model with $n = 1000$ and $m = 400$, as $m/n = 0.4$ on average for the Polis datasets, and the average number of voters is roughly 1000.

Conclusion 1: complete information algorithms. Our first question was whether GREEDY and LOCAL SEARCH-β achieve better CC-score in practice compared to two well-known committee election algorithms: APPROVAL VOTING and LOCAL SEARCH PAV.[11] To answer this question, we ran GREEDY (Algorithm 1), LOCAL SEARCH-β (Algorithm 3)[12], APPROVAL VOTING and LOCAL SEARCH PAV on the 118 data sets, running 20 random trials per dataset for both local search algorithms because of the random starting committee.

Figure 1 shows the CC-score attained by the four complete information algorithms on the 118 datasets. Since the synthetic data are drawn from the same distribution, we show the mean CC-score. This is not the case for the Polis data, which differ quite significantly, so that we plot the CC-score on each Polis dataset separately.

We can see that our GREEDY and LOCAL SEARCH algorithms achieve a higher (mean) CC-score than both APPROVAL VOTING and LOCAL SEARCH PAV across all Polis datasets, as well as on the synthetic data. Averaged across the Polis datasets, the best performing of our two algorithms achieves a CC-score 8.6% and 6.3% higher than LOCAL SEARCH PAV, APPROVAL VOTING, respectively. Across the synthetic data, this is 0.67%, 1.3%, respectively. We note that the CC-score on the synthetic data is generally very close to 1, leaving less room for improvement to begin with. At times APPROVAL VOTING outperforms LOCAL SEARCH PAV, which may be unexpected (e.g., datasets 3, 10, 15). We conclude from this that our two algorithms are significant improvements over these existing algorithms when the objective is to select diverse committees.

Conclusion 2: querying algorithms. To answer our second question, we take the CC-scores of GREEDY and LOCAL SEARCH from Figure 1 as our baseline and inspect how close we can get to these scores when information is incomplete and/or inaccurate, using a realistic number of queries. For the incomplete information model, taking $t = 20$, we write M for the expected number of query sets of size t that each voter is presented with. We then configure GREEDY-INCOMPLETE (Algorithm 2) and LS-INCOMPLETE (Algorithm 4) so that $M = 1, 2, 3, 4, 5$, which is much lower than what would theoretically be required. We do 50 random trials of each algorithm (due to the random querying of voters). For the inaccurate information model, we ignore the theoretically required repetitions of the queries, and run the standard greedy and local search algorithms, but with $p = 0.1$[13]. Finally, we combine both of the above by running Algorithm 2 and Algorithm 4 with the above values of M but for $p = 0.1$.

Figure 2 shows the CC-scores obtained in these experiments. We see that both algorithms with accurate responses ($p = 0$) obtain a score very close (0.85–0.95) to versions of the algorithms with complete information, even when $M = 1$. With complete information, both algorithms obtained an average CC-score of around 0.8. Multiplying by 0.9 still yields a score above the worst-case approximation ratio of $1 - 1/\mathrm{e} \approx 0.63$, even when the optimal solution would have a CC-score of 1. Example 8 highlights the great performance witnessed in the experiments for $p = 0$. This shows that even with limited querying of voters, we can reliably attain a diverse committee.

EXAMPLE 8. *Figure 2 shows that for $M = 5$, the CC-score attained by LS-INCOMPLETE, on average across 18 Polis datasets with 50 runs per set, is approximately 0.95 times the score attained by LOCAL SEARCH. The dataset* `vtaiwan.uberx` *has $m = 197$ and $n = 1921$. For these values of m and n, taking $\gamma = 0.95$, $k = 8$, $t = 20$, and $\delta = 0.05$, we would need $M = 7.109 \cdot 10^8$ to obtain the guaranteed ratio of $(1 - 1/e) \cdot 0.95$ with probability 0.95, using the upper bound proven in Theorem 5.*

With $p = 0.1$, but with complete information, the performance decreases by a factor of 0.95–0.97 on average, which yields a score confidently above the worst-case approximation ratio of $1 - 1/\mathrm{e}$. As a comparison: taking again dataset `vtaiwan.uberx` with $\delta = 0.05$ would require repeating each query 32 times to obtain the guaranteed ratio of $(1 - 1/\mathrm{e})$ with probability 0.95, according to the upper bound in Theorem 6.

Taking both $p = 0.1$ *and* incomplete information, not meeting the theoretically required number of queries for *either* scenario, the performance takes a noticeable hit. The algorithms perform worse than the versions with complete information by a factor of 0.7–0.8 for $M = 1$ increasing to 0.85–0.95 for $M = 5$. However, this still implies that we are above the worst-case approximation ratio. Do note that the standard deviation can become relatively large here, especially for the Polis data, which appear to be less homogeneous than the synthetic data.

[10] https://github.com/compdemocracy/openData.

[11] We configure LOCAL SEARCH PAV with $\alpha = 1$ which is the best performing configuration retaining a provably polynomial time runtime [15].

[12] We took β as in Theorem 5 and write LOCAL SEARCH going forward.

[13] We ran the experiments for multiple values of p and the results were consistent. We also ran the experiments for different values of k and the results were similar.

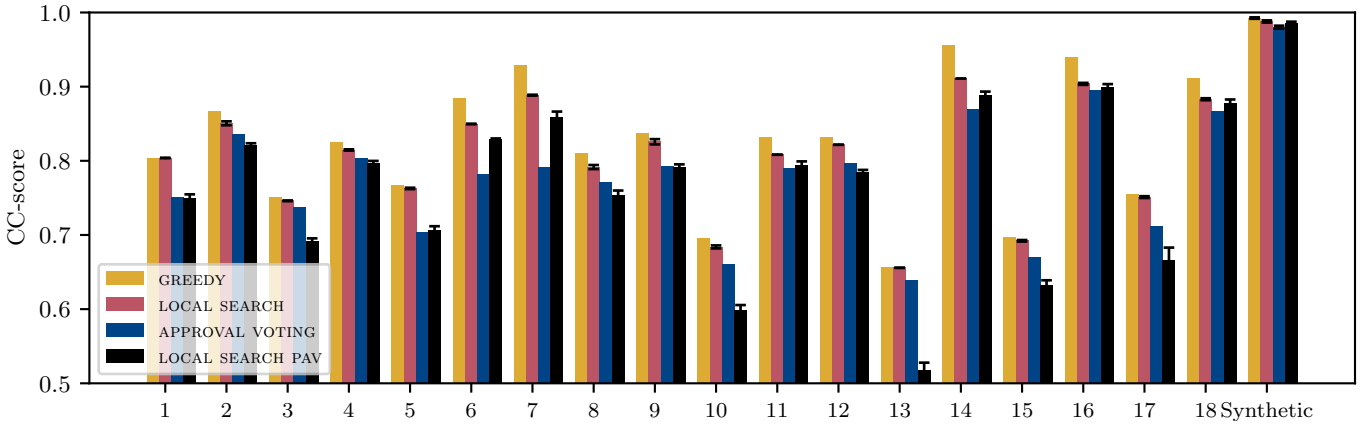

Figure 1: CC-score achieved by algorithms GREEDY, LOCAL SEARCH (Algorithms 1 and 3), APPROVAL VOTING and LOCAL SEARCH PAV on the 18 Polis data sets (plotted separately) and 100 synthetic data sets (plotted together) for $k = 8$. For the two local search algorithms, we ran 20 random trials (due to the random initial committee), plotting the mean and standard deviation. We used Paul Tol's high contrast colour scheme, designed to be color blind safe. See the full version of the paper for the names corresponding to the numbers of the datasets.

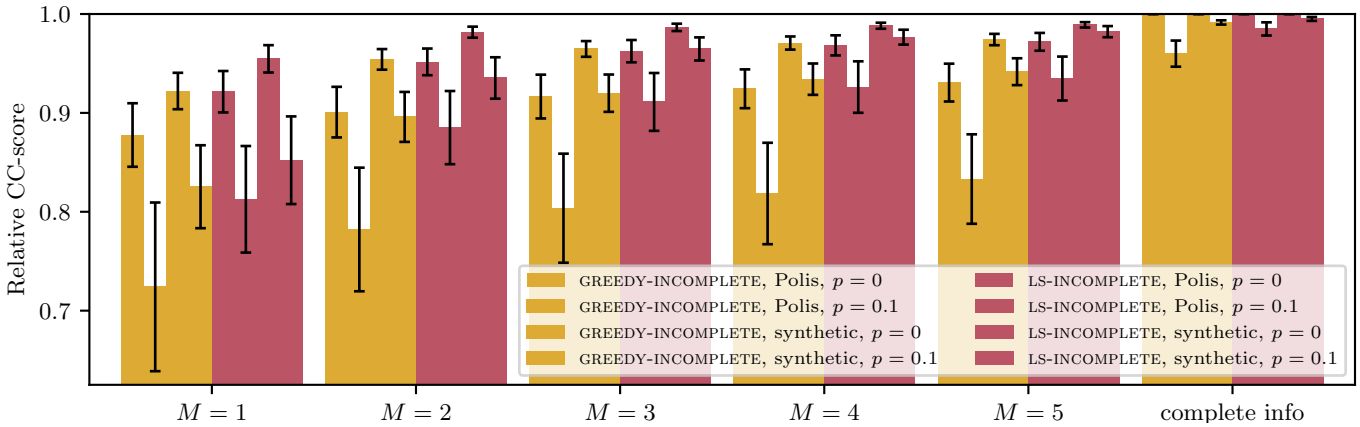

Figure 2: CC-score of GREEDY-INCOMPLETE (Algorithm 2) and LS-INCOMPLETE (Algorithm 4) on the 18 Polis and 100 synthetic data sets for $k = 8$. The algorithms are configured so that each voter is expected to be queried only $M = 1, 2, 3, 4, 5$ times. This was done for $p = 0$ (accurate responses) and $p = 0.1$ (inaccurate responses). On each of the 118 data sets, we ran 50 random trials of the algorithms, plotting the mean and standard deviation. All scores shown are relative to the complete and accurate information setting. We used Paul Tol's high contrast colour scheme, designed to be color blind safe.

7 OUTLOOK

Our work is the first to address diversity in approval-based committee elections in the context of online civic participation platforms. Measuring diversity by the Chamberlin-Courant score, we proved diverse committees can be found by querying only a small fraction of the voters, even when responses may be inaccurate. This remains true in the presence of external diversity constraints on the committee, such as quota. Our algorithms match lower bounds on the query complexity (up to log-factors). We verify these theoretical results empirically on both real-life and synthetic data.

Our results open up several directions for future research. *First*, it would be interesting to combine the incomplete and inaccurate information models explicitly in theoretical analysis. *Second*, it would be desirable to lift our results on inaccurate information to richer error models, e.g., where the error probability changes over time and/or between voters. *Third*, one could study more adaptive algorithms that can decide whom to best query about which comment at which time—the so-called 'comment routing problem' [35]. In doing so, the query complexity can perhaps be lowered further. *Fourth*, one can explore the significance of our results for active learning [33], with the aim of optimizing the querying of annotators in distributed data labelling tasks.

ACKNOWLEDGMENTS

We want to thank the anonymous reviewers of EC'25, COMSOC'25 and AAAI'26 for their helpful comments and suggestions. Feline,

Davide and Pradeep were supported by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://www.hybrid-intelligence-centre.nl/, under Grant No. (024.004.022). Davide ackowledges support by the European Union under the Horizon Europe project Perycles (Participatory Democracy that Scales, https://perycles-project.eu/).

This project was funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.


Funded by
the European Union

REFERENCES

[1] Haris Aziz, Markus Brill, Vincent Conitzer, Edith Elkind, Rupert Freeman, and Toby Walsh. 2017. Justified Representation in Approval-Based Committee Voting. *Social Choice and Welfare* 48, 2 (2017), 461–485.

[2] Jan Behrens, Axel Kistner, Andreas Nitsche, and Björn Swierczek. 2014. *The Principles of Liquid Feedback.* Interaktive Demokratie.

[3] Craig Boutilier and Jeffrey S Rosenschein. 2016. *Handbook of computational social choice.* Cambridge University Press, Chapter Incomplete Information and Communication in Voting, 223–257.

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

[5] Markus Brill. 2018. Interactive Democracy. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS).* 1183–1187.

[6] Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. 2011. Maximizing a Monotone Submodular Function Subject to a Matroid Constraint. *SIAM J. Comput.* 40, 6 (2011), 1740–1766. https://doi.org/10.1137/080733991

[7] John R. Chamberlin and Paul N. Courant. 1983. Representative Deliberations and Representative Decisions: Proportional Representation and the Borda Rule. *American Political Science Review* 77, 3 (1983), 718–733. https://doi.org/10.2307/1957270

[8] Vincent Cohen-Addad, Anupam Gupta, Amit Kumar, Euiwoong Lee, and Jason Li. 2019. Tight FPT Approximations for k-Median and k-Means. In *46th International Colloquium on Automata, Languages, and Programming, ICALP (LIPIcs, Vol. 132)*, Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi (Eds.). 42:1–42:14. https://doi.org/10.4230/LIPICS.ICALP.2019.42

[9] Christina Eder, Ingvill Constanze Mochmann, and Markus Quandt. 2015. *Political Trust and Disenchantment with Politics: International Perspectives.* Number volume 125 in International Studies in Sociology and Social Anthropology. Brill.

[10] Piotr Faliszewski, Piotr Skowron, Arkadii Slinko, and Nimrod Talmon. 2017. *Trends in Computational Social Choice.* AI Access Foundation, Chapter Multiwinner Voting: A New Challenge for Social Choice Theory, 27–47.

[11] Uriel Feige. 1998. A Threshold of ln(n) for Approximating Set Cover. *Journal of the Association for Computing Machinery (ACM)* 45, 4 (1998), 634–652. https://doi.org/10.1145/285055.285059

[12] Yuval Filmus and Justin Ward. 2013. The Power of Local Search: Maximum Coverage over a Matroid. In *Proceedings of the 29th Symposium on Theoretical Aspects of Computer Science (STACS) (LIPIcs, Vol. 14).* 601–612. https://doi.org/10.4230/LIPICS.STACS.2012.601

[13] Sara Fish, Paul Gölz, David C Parkes, Ariel D Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. 2024. Generative Social Choice. In *Proceedings of the 25th ACM Conference on Economics and Computation.* 985–985.

[14] Davide Grossi, Ulrike Hahn, Michael Mäs, Andreas Nitsche, Jan Behrens, Niclas Boehmer, Markus Brill, Ulle Endriss, Umberto Grandi, Adrian Haret, Jobst Heitzig, Nicolien Janssens, Catholijn M. Jonker, Marijn A. Keijzer, Axel Kistner, Martin Lackner, Alexandra Lieben, Anna Mikhaylovskaya, Pradeep K. Murukannaiah, Carlo Proietti, et al. 2024. Enabling the Digital Democratic Revival: A Research Program for Digital Democracy. (2024). arXiv preprint arXiv:2401.16863.

[15] Daniel Halpern, Gregory Kehne, Ariel D. Procaccia, Jamie Tucker-Foltz, and Manuel Wüthrich. 2023. Representation with Incomplete Votes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 5657–5664. https://doi.org/10.1609/aaai.v37i5.25702

[16] Dorit S. Hochbaum and Anu Pathria. 1998. Analysis of the Greedy Approach in Problems of Maximum k-Coverage. *Naval Research Logistics* 45, 6 (1998), 615–627. https://doi.org/10.1002/(SICI)1520-6750(199809)45:6<615::AID-NAV5>3.0.CO;2-5

[17] Aviram Imber, Jonas Israel, Markus Brill, and Benny Kimelfeld. 2022. Approval-Based Committee Voting under Incomplete Information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5076–5083.

[18] Russell Impagliazzo and Ramamohan Paturi. 2001. On the Complexity of k-SAT. *J. Comput. System Sci.* 62, 2 (2001), 367–375. https://doi.org/10.1006/jcss.2000.1727

[19] Michael J Kearns and Umesh Vazirani. 1994. *An Introduction to Computational Learning Theory.* MIT press.

[20] Martin Lackner and Piotr Skowron. 2023. *Multi-Winner Voting with Approval Preferences.* Springer International Publishing. https://doi.org/10.1007/978-3-031-09016-5

[21] Hélène Landemore. 2017. *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many* (first paperback printing ed.). Princeton University Press.

[22] Zihan Li, Pasin Manurangsi, Jonathan Scarlett, and Warut Suksompong. 2024. Complexity of Round-Robin Allocation with Potentially Noisy Queries. In *17th International Symposium on Algorithmic Game Theory*, Guido Schäfer and Carmine Ventre (Eds.), Vol. 15156. Springer Nature Switzerland, 520–537. https://doi.org/10.1007/978-3-031-71033-9_29

[23] Pasin Manurangsi. 2020. Tight Running Time Lower Bounds for Strong Inapproximability of Maximum k-Coverage, Unique Set Cover and Related Problems (via t-Wise Agreement Testing Theorem). In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA*, Shuchi Chawla (Ed.). SIAM, 62–81. https://doi.org/10.1137/1.9781611975994.5

[24] Tomáš Masařík, Grzegorz Pierczyński, and Piotr Skowron. 2024. A Generalised Theory of Proportionality in Collective Decision Making. In *Proceedings of the 25th ACM Conference on Economics and Computation.* Association for Computing Machinery, 734–754. https://doi.org/10.1145/3670865.3673619

[25] John G Matsusaka. 2020. *Let the People Rule: How Direct Democracy Can Meet the Populist Challenge.* Princeton University Press.

[26] Anna Mikhaylovskaya. 2024. Enhancing Deliberation with Digital Democratic Innovations. *Philosophy & Technology* 37, 1 (2024), 3.

[27] Anna Mikhaylovskaya and Élise Rouméas. 2024. Building Trust with Digital Democratic Innovations. *Ethics and Information Technology* 26, 1 (2024), 1. https://doi.org/10.1007/s10676-023-09736-4

[28] George Lann Nemhauser, Laurence Alexander Wolsey, and Marshall. L. Fisher. 1978. An Analysis of Approximations for Maximizing Submodular Set Functions—I. *Mathematical Programming* 14, 1 (1978), 265–294. https://doi.org/10.1007/BF01588971

[29] James B. Orlin, Abraham P. Punnen, and Andreas S. Schulz. 2004. Approximate Local Search in Combinatorial Optimization. *SIAM J. Comput.* 33, 5 (2004), 1201–1214. https://doi.org/10.1137/S0097539703431007

[30] Dominik Peters. 2018. Single-Peakedness and Total Unimodularity: New Polynomial-Time Algorithms for Multi-Winner Elections. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

[31] Dominik Peters and Martin Lackner. 2020. Preferences Single-Peaked on a Circle. *Journal of Artificial Intelligence Research* 68 (2020), 463–502.

[32] Manon Revel, Smitha Milli, Tyler Lu, Jamelle Watson-Daniels, and Maximilian Nickel. 2025. Representative Ranking for Deliberation in the Public Sphere. In *Forty-second International Conference on Machine Learning.*

[33] Burr Settles. 2009. *Active Learning Literature Survey.* Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

[34] Piotr Skowron and Piotr Faliszewski. 2015. Fully Proportional Representation with Approval Ballots: Approximating the MaxCover Problem with Bounded Frequencies in FPT Time. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29. https://doi.org/10.1609/aaai.v29i1.9432

[35] Christopher Small, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. 2021. Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. *Recerca: Revista de Pensament i Anàlisi* 26, 2 (2021).

[36] Krzysztof Sornat, Virginia Vassilevska Williams, and Yinzhan Xu. 2022. Near-Tight Algorithms for the Chamberlin-Courant and Thiele Voting Rules. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI 2022).* International Joint Conference on Artificial Intelligence, 482–488. https://doi.org/10.48550/arXiv.2212.14173

[37] Stanisław Szufa, Piotr Faliszewski, Lukasz Janeczko, Martin Lackner, Arkadii Slinko, Krzysztof Sornat, and Nimrod Talmon. 2022. How to Sample Approval Elections?. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 496–502. https://doi.org/10.24963/ijcai.2022/71

[38] Luis Sánchez-Fernández, Edith Elkind, Martin Lackner, Norberto Fernández, Jesús Fisteus, Pablo Basanta Val, and Piotr Skowron. 2017. Proportional Justified Representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31. https://doi.org/10.1609/aaai.v31i1.10611