Diverse committees with incomplete or inaccurate approval ballots

Feline Lindeboom, Martijn Brehm, Davide Grossi and Pradeep Murukannaiah COMSOC 2025, 17-19 September

Setting online deliberation platforms where participants input statements and (dis)agree with statements provided by their peers

Make a summary of k statements based on (dis)agreements

Elect a committee of k based on approval votes

n voters

m candidates

k committee members

Aim: elect a diverse committee

Chamberlin-Courant ⟷ Max Cover

Algorithms

The **Max Cover problem is NP-complete**. A `greedy` and a `local search` algorithm attain the best possible approximation ratio of 1-1/e (assuming P ≠ NP).

`greedy`: start from an empty committee and in each iteration add the candidate that yields the largest immediate gain in score

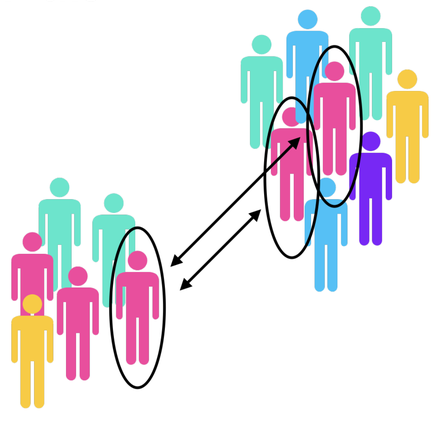
`local search`: start from an arbitrary committee and swap candidates until the increase in score is small enough

For the generalized problem of Max Cover over a *matroid constraint*!

This lets us implement quota

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

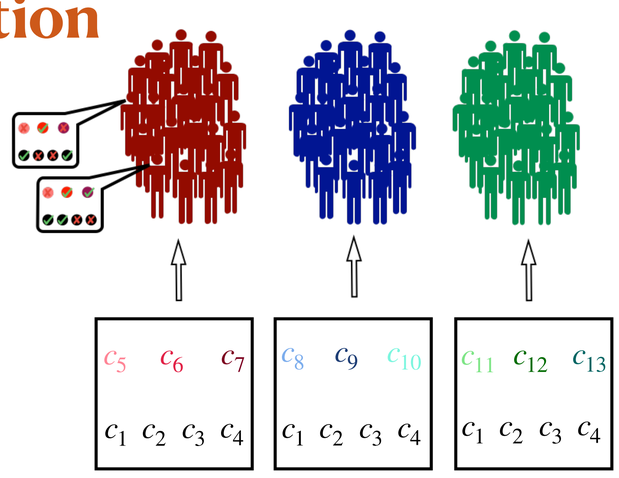
Table 1: Overview of query complexity bounds in m and n.



Incomplete information

Voters won't respond to all statements

Idea: per statement, query *part of* the voters and estimate the total population's opinion.



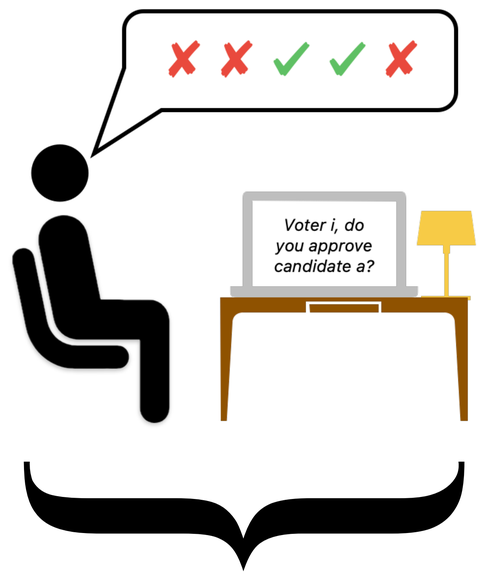
Lower bound (non-adaptive queries) Any algorithm needs to make $\Omega(m^2)$ non-adaptive queries to get arbitrarily close to a $(1-1/e)$ -approx. Max Cover score w.h.p.

Upper and lower bound (adaptive queries) Any algorithm needs to make $\Omega(m)$ adaptive queries to get arbitrarily close to a $(1-1/e)$ -approx. Max Cover score w.h.p.. Our `greedy` and `local search` algorithms both match this (up to log-factors), having query complexity $O(m \log m) = \tilde{O}(m)$.

Inaccurate information

Voters may be inaccurate: approvals are corrupted independently with a fixed probability $p \in (0,1/2)$.

Idea: repeatedly query voters the same question, and interpret the majority winner as their 'true' answer



Voter i doesn't approve candidate a

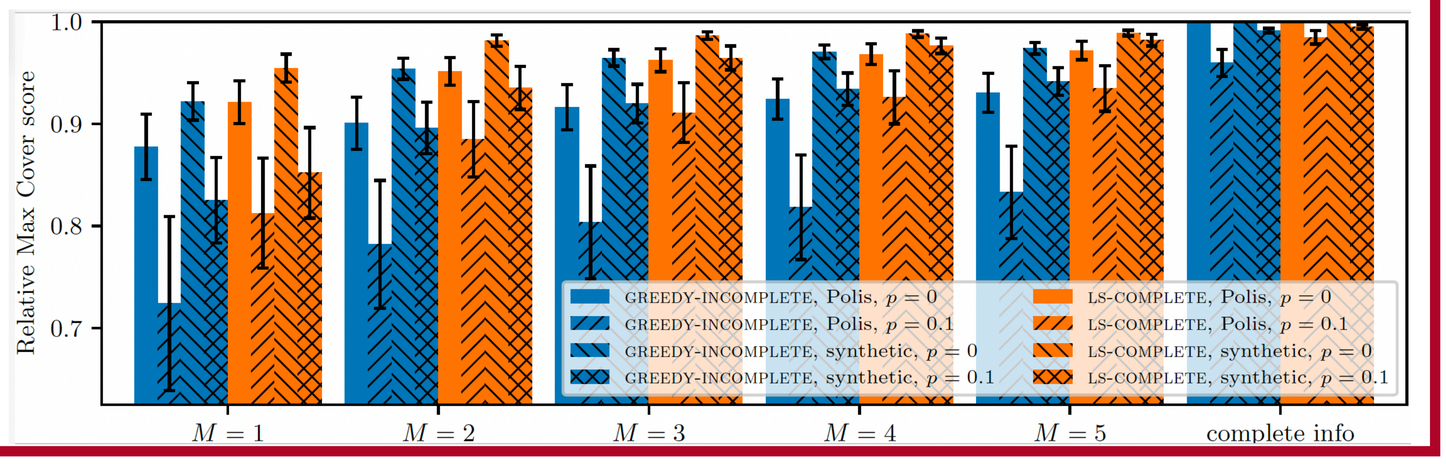
Lower bound Any algorithm needs to make $\Omega(nm \log(1/\delta))$ p-inaccurate queries to be $(1-1/e)$-approx. for Max Cover w.p. $1-\delta$.

Upper bound There is an algorithm that makes $\mathcal{O}(nm \log(nm/\delta))$ p-inaccurate queries and is $(1-1/e)$-approx. for Max Cover w.p. $1-\delta$.

Experiments

Our algorithms beat the optimal approx. ratio of $(1-1/e)$, even when voters are queried a small constant number of times (vs $\tilde{O}(m)$ times) , even when information is inaccurate *and* incomplete.

We studied 18 Polis datasets and 100 datasets sampled from the (0.0891,0.693)-resampling model (Szufa et al, 2022)[1], which is structurally similar to Polis data.

1 https://arxiv.org/abs/2207.01140

Example dataset: `vtaiwan.uberx`. $m = 197, n = 1921$. Theory tells us attaining approx. ratio $0.95(1-1/e)$ w.p. at least 0.95 requires $1.422 \cdot 10^9$ queries p.p.