# Efficient Vote Elicitation under Candidate Uncertainty

**Craig Boutilier** and **Yuval Filmus** and **Joel Oren**
Department of Computer Science, University of Toronto

## Abstract

Top-$k$ voting is an especially natural form of partial vote elicitation in which only length $k$ prefixes of rankings are elicited. We analyze the ability of top-$k$ vote elicitation to correctly determine true winners, with high probability, given probabilistic models of voter preferences and candidate availability. We provide bounds on the minimal value of $k$ required to determine the correct winner under the plurality and Borda voting rules in both the worst-case preference profiles and under impartial culture and Mallows models; and we derive conditions under which the special case of *zero-elicitation* (i.e., $k = 0$) produces the correct winner. Empirical results confirm the value of top-$k$ voting.

## 1 Introduction

Social choice has provided valuable foundations for the development of computational approaches to preference aggregation, group decision making and a variety of other problems in recent years. As algorithmic advances and data accessibility make the methods of social choice more broadly applicable, relaxing the assumptions of classical models to fit a richer class of practical problems becomes imperative. To this end, research has begun to address the *informational demands* of preference aggregation, considering models in which information about, say, available candidates [13; 1; 3], or voter preferences [11; 22; 15] may be incomplete.

In this work, we bring together two such lines of research to investigate the feasibility and value of *top-$k$ voting*. Our first main motivation is to use intelligent *vote elicitation* techniques to minimize the amount of voter preference information required to determine the winner in an election (or more broadly, the desired outcome of a group decision). Vote elicitation has received considerable attention recently [7; 6; 10; 15; 16; 8], and has proven to be effective in reducing the amount of information, and corresponding cognitive and communication burden, needed to determine winners in practice. Our second motivation is handling uncertainty in the set of available candidates. In many settings, voters may need to specify their preferences over a range of potential candidates prior to knowing which are in fact available or viable for selection [13; 1; 3]. Examples include ranking job candidates, public projects, or even restaurants. The potential impact of candidate unavailability on vote elicitation is clear: since certain desirable alternatives may turn out to be unavailable, one may need to elicit more preference information than is typical in the case of fully known candidates in order to ensure the correct winner is chosen.

We address efficient preference elicitation in this context in the form of *top-$k$ elicitation*, in which the agents are asked to provide the length $k$ prefix of their ranking instead of their full ranking. In the standard "known candidates" model, top-$k$ voting has been used heuristically [10] and the optimal choice of $k$ has been analyzed from a sample-complexity-theoretic perspective [16]. However, bounds on the required values of $k$ for specific preference distributions and voting rules have remained unaddressed, as has the impact of unavailable candidates on top-$k$ voting. We focus on two common voting rules, plurality and Borda. Given a prior distribution on the preference profile, and a distribution over the set of available candidates (for which the standard "known candidates" model is a special case), we ask: what is the minimal value of $k$ for which top-$k$ voting determines the true winner (with high probability), with respect to the underlying preference profile? We provide theoretical results, in the form of upper and lower bounds on $k$, for both worst-case preferences and certain preference distributions (including impartial culture and Mallows distributions). As a special case, we consider *zero-elicitation protocols*, where $k = 0$: as a function of election parameters, we show when the true winner can be determined with high probability without eliciting *any* information from voters. We also provide empirical results demonstrating the extent to which top-$k$ voting determines true winners as a function of $k$.

## 2 The Model

Let $C = \{c_1, \ldots, c_m\}$ be the set of (potential) candidates from which a winner is to be selected using some voting rule. Let $N = \{1, \ldots, n\}$ be the set of voters, and let voter $i$'s *preference* $\pi_i$ be a permutation of $C$: intuitively, for $1 \leq j < j' \leq m$, $\pi_i(j)$ is preferred by $i$ to $\pi_i(j')$. Let $\mathcal{L}$ denote the set of all preferences over $C$. A *preference profile* $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n) \in \mathcal{L}^n$ represents the collection of voter preferences. A *voting rule* $v \colon \mathcal{L}^n \times 2^C \to C$ selects a winner from $C$ given a vote profile and a set of available candidates.

We consider two voting rules: plurality and Borda. In *plurality voting*, given a profile $\boldsymbol{\pi}$, the plurality score of

candidate $c$ is the number of times that $c$ is ranked first: $sc^P(c, \boldsymbol{\pi}) = |\{i \in N : \pi_i(1) = c\}|$. The plurality winner is the candidate with maximal plurality score (ties can be handled arbitrarily; the tie-breaking rule used does not impact our results). In *Borda* voting, the *Borda score* of $c$ is the number of candidates ranked below it, summed over all preferences $\pi_i$: $sc^B(c, \boldsymbol{\pi}) = \sum_{i \in N}[m - \pi_i^{-1}(c)]$. The Borda winner is the candidate with maximal Borda score.

**Unavailable Candidates.** Recent attention has been paid to the possibility of voting over a slate of *potential candidates* $C$, prior to determining the availability of the *actual set* of candidates $A \subseteq C$. When determining the availability of candidates is costly or risky (e.g., making job offers, determining feasibility of public projects, calling restaurants for reservations), it often makes sense to elicit voter preferences prior to determining availability, focusing availability determination on candidates most likely to be winners relative to the true available set $A$ [13; 1; 3]. Following these recent models, we assume that each candidate $c \in C$ is available i.i.d. with some fixed probability $p \in (0, 1]$ (this simplifies our presentation, but having distinct availabilities $p_c$ for different $c$ does not change the nature of our results, which can be adapted accordingly).

Given a set $A \subseteq C$ of available candidates, a *reduced preference* $\pi_i|_A$ is obtained by restricting $\pi_i$ to the candidates in $A$; we denote by $\boldsymbol{\pi}|_A$ the *reduced preference profile* obtained in this way. Plurality and Borda voting in the unavailable candidate model are determined in a straightforward way, using the scores obtained relative to the reduced profile. Notice that in the unavailable candidates model, it is no longer sufficient to run plurality voting by eliciting just the top-ranked candidate from each voter: in general the entire ranking is needed.

**Top-$k$ voting.** Recent research has focused on the use of intelligent preference elicitation schemes to minimize the burden on voters and obviate the need to provide full preference rankings. One especially natural approach is *top-$k$* voting, in which voters are asked to list only their $k$ most preferred candidates (or the $k$-th prefix of their ranking) [10; 15; 16]. We discuss below alternative ways in which such votes can be used to determine winners; but here we adopt an especially simple approach.

Given a voting rule $v$ and some $k \in [m]$, we denote by $(\boldsymbol{\pi}^{(k)}) = (\pi_1^{(k)}, \ldots, \pi_n^{(k)})$ the $k$-truncated preference profile. We use this truncated profile to determine plurality scores in the obvious fashion, by counting the number of first place rankings. We compute Borda scores by assigning a score of $\tilde{m} - \pi_i^{-1}(c)$ to any candidate $c$ in voter $i$'s $k$-truncated vote, where $\tilde{m}$ is the number of available candidates, and a score of zero otherwise. In the unavailable candidates model, we employ the same technique, restricting the *truncated* vote to the available set $A$. Our goal is to determine values of $k$ that suffice to determine the true winner (with high probability) relative to the true (untruncated) preference profile.

If candidates are always available (i.e., $p = 1$) then $k = 1$ is sufficient to determine the correct plurality winner, and general top-$k$ voting is of no value. In contrast, the possibility of unavailable candidates intuitively requires that one use larger values of $k$ for most voting rules.

**Probabilistic preference models.** It has become increasingly common to analyze voting rules under the assumption that agent preferences are drawn from a prior distribution over permutations. One important class of distributions, widely used in psychometrics, statistics, and machine learning, is the *Mallows $\varphi$-distribution* [17; 18]. It is described by two parameters: a *reference ranking* $\hat{\pi} \in \mathcal{L}$, and a *dispersion parameter* $\varphi$ (controlling variance). The probability of a permutation $\pi$ under this model is $\Pr(\pi) = \varphi^{\tau(\pi,\hat{\pi})}/Z_m$, where $\tau(\pi, \hat{\pi})$ is the Kendall-tau distance,

$$\tau(\pi_1, \pi_2) = |\{c, c' : \pi_1^{-1}(c) < \pi_1^{-1}(c') \text{ and } \pi_2^{-1}(c) > \pi_2^{-1}(c')\}|,$$

and $Z_m$ is a normalization constant. Importantly, when $\varphi = 1$, one obtains the uniform distribution over $\mathcal{L}$, the so-called *impartial culture (IC)* assumption, a modeling assumption widely used in social choice.

**Related Work.** As mentioned, vote elicitation has attracted considerable recent attention, usually in the context of standard "known available candidate" models. Of particular relevance is work on top-$k$ voting. Unlike our model, in which we "zero out" the scores of unavailable candidates, other work has treated the uncertainty in the missing candidates more cautiously. Kalech et al. [10] use top-$k$ ballots to determine possible and necessary winners [11] and develop heuristic elicitation schemes to extend these ballots to quickly identify true winners for several different voting rules. Lu and Boutilier [15] use *minimax regret* to measure error in winner determination and to guide elicitation heuristically as well. Both methods show good empirical performance (and handle general partial votes) but provide no theoretical guarantees on the required values of $k$. The optimal choice of $k$ has been analyzed from a sample-complexity-theoretic perspective in [16], which provides bounds on the *required number of sampled profiles* needed to estimate the required value of $k$ for arbitrary distributions; but this does not provide direct bounds on $k$ itself. None of these models considers unavailability.

The idea of voting with unavailable candidates was considered in [13; 1], who study the impact of missing candidates on the fidelity of a winner using voting rules such as Borda, and how close *ranking policies* for selecting winners approximate the true winner. More general querying policies, assuming costly availability tests, were studied in [3]. Unavailable candidate models also bear a strong connection to the study of manipulation by candidate addition and deletion [9; 2]. These models do not consider partial preferences. Chevalyre et al. [5] consider the possible and necessary winner problem under (general) partial preferences, when new candidates are added to an election, for several voting rules, but do not consider elicitation or quantifying the amount of information needed to determine a necessary winner.

**Our results.** In most of our theoretical bounds, we say that a value of $k$ *produces a correct winner with high probability (w.h.p.)* if the probability that top-$k$ voting returns the true (full profile) winner is $1 - o(1)$, where $o(1) \to 0$ as $m \to \infty$. For plurality, we provide an upper bound of $O(\log m)$ on the $k$ that produces the correct winner w.h.p., if $n$ is polynomial in $m$, even if the preference profile is selected by an adver-

| Voting rule | Adv. preferences | IC |
|---|---|---|
| Plurality, $n = poly(m)$ | $k = O(\log m)$ | $k = O(\log m)$ |
| Plurality, $n = \exp(m)$ | $k = \Omega(m)$ | $k = \Theta(\log m)$ |
| Borda, $n = \Omega(m^3 \log m)$ | $k = \Omega(m)$ | $k = \Omega(m/\log m)$ |

**Table 1**: Top-$k$ voting: bounds on $k$

sary. If $n$ is exponentially larger than $m$, we show that under impartial culture we require $k = \Theta(\log m)$, while $k = \Omega(m)$ is needed in the worst case. For Borda, we show that for a sufficiently large $n$ (polynomial in $m$), $k$ is $\Omega(m/\log m)$ under impartial culture, even if $p = 1$; and it has a lower-bound of $k = \Omega(m)$ in the worst case. Our top-$k$ results are summarized in Table 1.

We also provide theoretical results for the special case of $k = 0$, or *zero elicitation*, and for cases where preferences are distributed according to a Mallows model with reference ranking $\hat{\pi}$, providing lower bounds on the required number of voters $n$ needed to find winners w.h.p., as a function of $\varphi$ and $m$. For plurality, we show that if $n = \Omega(\log m/(1-\varphi)^3)$, then the top candidate in $\hat{\pi}$ is the winner w.h.p. For Borda, we derive a lower bound of $\ln m \cdot \Gamma(\varphi)$ on $n$, where $\Gamma(\varphi) = (8(1+\varphi)^2(1-\varphi)^3 + (1+\varphi))/(1-\varphi)^7$.

We support our theoretical findings by testing the performance of top-$k$ voting (including the special case of zero elicitation) under varying parameter values $(k, n, m, \varphi)$. Our empirical results suggest that when the dispersion parameter is bounded away from 1, fairly low values of $k$ are sufficient for correct winner determination.

Space precludes inclusion of proofs for all results. Omitted proofs can be found in a longer version of this paper (this will be made available online after the reviewing period).

## 3 Top-$k$ Voting and Plurality Scoring

We start with a theoretical analysis of the performance of top-$k$ voting with plurality scoring, assessing the values of $k$ needed to determine the true plurality winner w.h.p. As noted above, if the candidate availability probability $p$ is 1, setting $k = 1$ trivially guarantees correct winner selection. Therefore, in this section we assume that $p$ is a *fixed* probability, bounded away from 1. We distinguish: (a) *worst-case results*, in which an adversarial preference profile is selected to minimize the odds of correct winner selection, and expectations are taken over available sets $A$; and (b) *average-case results*, in which profiles are drawn from some distribution (e.g., impartial culture), and expectations are taken over both profiles and available sets.

We first show that, even in the worst case, when the number of voters $n$ is "small" relative to the number of candidates $m$, a small value of $k$ suffices for plurality:

**Theorem 1** (Worst-case upper bound, poly. $n$)**.** *If $n = poly(m)$, then top-$k$ voting with $k = O(\log m)$ determines the correct plurality winner w.h.p. in the worst case.*

*Proof.* Consider a vote $\pi \in \mathcal{L}$. Set $k = 2\log n/\log(\frac{1}{1-p})$. The probability that all top-$k$ candidates are unavailable is $1/n^2$. Taking a union bound over all votes, the probability that some vote has all top-$k$ candidates unavailable is $1 - 1/n = 1 - o(1)$. □

Since this $O(\log m)$ upper bound applies in the worst case, it also applies to the average case for any profile distribution. However, in the worst case, having $n$ sub-exponential in $m$ is required if we want a small $k$.

**Theorem 2** (Worst-case lower bound, exp. $n$)**.** *If $n = \exp(poly(m))$, top-$k$ voting requires $k = \Omega(m)$ to determine the correct plurality winner w.h.p. in the worst-case.*

*Proof.* Let $C = \{c_1, \ldots, c_m\} \cup \{a, b\}$, and $p = 1/2$. A key observation is that the unavailable set has size at least $m/2$ with probability very close to $1/2$ (we assume for simplicity that $m$ is even). We create a scenario in which $a$ and $b$ have very close plurality scores, requiring a large value of $k$ to tell which has the higher score. Consider the set $\mathcal{H} = \{S \subseteq C : |S| = m/2\}$ containing all subsets of $C$ of size $m/2$. We show that $k \geq m/2$ is required. Create two sets of votes:

1. $V_1$: This set ensures $a$ and $b$ have the two highest scores if available (which occurs with constant probability, so assume both are). Let $t = 2 \cdot |\mathcal{H}|$, and for a set $S \subseteq C$, let $lin(S)$ be an arbitrary ordering of $S$. Create $t+1$ copies of $a \succ lin(C \setminus \{a\})$, and $t$ copies of $b \succ lin(C \setminus b)$. Note: $a$ gets one more vote than $b$ in $V_1$.
2. $V_2$: For every $S \in \mathcal{H}$, create two copies of the ranking $lin(S) \succ b \succ a \succ lin(C \setminus (S \cup \{a, b\}))$.

Now, suppose the unavailable set has size at least $m/2$. The plurality score of $a$ is $t+1$, the score of $b$ is at least $t+2$, and so $b$ is the true winner. Otherwise, the score of $a$ is $t+1$, that of $b$ is $t$, and $a$ is the winner. (All other candidates have score at most $t$.) If $k \leq m/2$ then the voting scheme doesn't see $b$ in the set $V_2$, and so it gives incorrect results with probability roughly $p^2/2$. □

Thus, for large $n$, we must set $k \geq m/2$ in the worst-case. However, under impartial culture, a small value of $k = O(\log m)$ again suffices for the average case:

**Theorem 3** (Avg. case upper bound, exp. $n$)**.** *If $n = \exp(\Omega(m))$, then top-$k$ voting with $k = O(\log m)$ determines the correct plurality winner w.h.p. under impartial culture.*

*Proof.* Partition $V$ into two sets: $V_1 = \{\pi_i \in V :$ one of $\pi_i(1), \ldots, \pi_i(k)$ is available$\}$, $V_2 = V \setminus V_1$. Let $A \subseteq C$ be the available set, let $\tilde{m} = |A|, n_1 = |V_1|, n_2 = |V_2|$. For $c \in C$, let $sc_1^P(c)$ and $sc_2^P(c)$ be its plurality scores in elections $(V_1, A), (V_2, A)$, respectively. W.l.o.g., order candidates based on $sc_1^P(\cdot)$: $sc_1^P(c_1) \geq sc_1^P(c_2) \geq \cdots \geq sc_1^P(c_{\tilde{m}})$. We prove that $c_1$ is the true winner w.h.p.

By a simple Chernoff-bound argument, $\frac{m \cdot p}{2} \leq \tilde{m} \leq 2m \cdot p$, w.h.p. Similarly, a simple calculation shows that $\mathbb{E}[n_2] = n \cdot (1-p)^k$, and using a Chernoff bound we obtain $n_2 \leq 2n \cdot (1-p)^k$ w.h.p. Hence, $n_1 \geq n - 2n \cdot (1-p)^k$ w.h.p.

We now give an anti-concentration argument about the difference between the scores according to $V_1$. We let $D_{i,j}^1 = sc_1^P(c_i) - sc_1^P(c_j)$ (we define $D_{i,j}^2$ similarly).

**Lemma 4.** *$D_{1,2}^1 = \Omega(n_1/m^{3.5})$ with high probability.*

*Proof.* After conditioning on $A$, consider the votes $V_1$ sequentially. By a simple balls and bins argument, the difference between the scores of $c_i$ and $c_j$ increases by 1 due to

vote $\pi_t$ ($t = 1, \ldots, n_1$) with probability $1/\tilde{m}$, decreases by 1 with probability $1/\tilde{m}$, and does not change with probability $1 - 2/\tilde{m}$. We can thus treat this change as a random variable $X_t$, rewriting $D^1_{i,j} = \sum_{t=1}^{n_1} X_t$, where $X_t = 1, X_t = -1$ each with probability $1/\tilde{m}$, and $X_t = 0$ with probability $1 - 2/\tilde{m}$. Then $Var(X_t) = \mathbb{E}[X_i^2] = \frac{2}{\tilde{m}}, \mathbb{E}[D^1_{i,j}] = \frac{2n_1}{\tilde{m}}$, and $\rho = \mathbb{E}[|X_t|^3] = \frac{2}{\tilde{m}}$. The Berry-Esseen Theorem allows us to prove that $D^1_{1,2}$ (and hence $D^1_{1,j}$ for every $j$ s.t. $c_j \in A$) is "large enough."

**Lemma 5** (Berry-Esseen [12])**.** *Let $X = X_1 + \cdots + X_n$ be the sum of i.i.d. zero-mean random variables s.t. $\mathbb{E}[X_i^2] = \sigma^2 > 0, \mathbb{E}[|X_i|^3] = \rho < \infty$. Let $F_n(\cdot)$ be the cdf of $X$, and let $\Phi(\cdot)$ be the cdf of the normal distribution. Then:*

$$\sup_x |F_n(x) - \Phi(x)| < \frac{C\rho}{\sigma^3\sqrt{n}} \tag{1}$$

*where $0 < C \leq 0.4784$.*

In our case: $\frac{C\rho}{\sigma^3\sqrt{n}} = \frac{C}{\sqrt{n_1}} \cdot \frac{2}{\tilde{m}} \cdot \left(\frac{\tilde{m}}{2}\right)^{3/2} = C' \cdot \sqrt{\frac{\tilde{m}}{n_1}}$. Hence, we may assume that $D^1_{i,j}$ is effectively given by the normal distribution $\mathcal{N}(0, \sigma^2 = \frac{2n_1}{\tilde{m}})$, which gives us:

$$\begin{aligned}\Pr[|D^1_{i,j}| < t] &\leq \frac{1}{\sqrt{2\pi}} \int_{-t/\sigma}^{t/\sigma} e^{-x^2/2} \mathrm{d}x < \frac{1}{\sqrt{2\pi}} \cdot (2t/\sigma) \\ &= t \cdot \sqrt{\frac{\tilde{m}}{\pi \cdot n_1}}\end{aligned} \tag{2}$$

Setting $t = \frac{\sqrt{n_1}}{\tilde{m}^{3.5}}$ and taking the union bound over all possible pairs $(i, j)$ gives $D^1_{1,2} = \Omega(\frac{n_1}{\tilde{m}^{3.5}})$ with probability $1 - O(1/\tilde{m}) = 1 - o(1)$, where the last equality follows from the concentration bound on $\tilde{m}$. □

A concentration bound on $D^2_{1,j}$ (for all $c_j \in A \setminus \{c_1\}$) follows from a Chernoff bound and a union bound over all $j$:

$$\Pr[D^2_{1,j} \leq 2\sqrt{\frac{n_2}{\tilde{m}}} \cdot \log m \text{ for all } j \geq 2] = 1 - o(1) \tag{3}$$

We now summarize by showing that, w.h.p., $D^1_{1,2} > D^2_{1,j}$:

$$\frac{\sqrt{n_1}}{\tilde{m}^{3.5}} > \frac{\sqrt{n_2} \cdot \log m}{\sqrt{\tilde{m}}} \tag{4}$$

As $m > \tilde{m}$ and $\sqrt{m} > 1$, it suffices to show:

$$\frac{\sqrt{n - 2n \cdot (1-p)^k}}{m^{3.5}} > \sqrt{2n \cdot (1-p)^k} \cdot \log m \tag{5}$$

The above holds (for $n, m$ sufficiently large) if we set $(1 - p)^k = m^{-8}$, which gives $k = O(\log m)$, as required. □

A matching lower-bound shows this upper-bound is tight:

**Theorem 6** (Avg. case lower bound)**.** *If $n = \exp(\Omega(m))$, $k = \Omega(\log m)$ is necessary for top-$k$ voting to produce the true plurality winner w.h.p. under impartial culture.*

*Proof.* The proof is largely symmetric to the proof of the upper-bound. We use the same notation as in the previous proof. We first prove an upper-bound on the difference between the score of the highest-ranking candidate and the second-highest. As before, order $C$ based on their scores in $V_1$: $sc^P_1(c_1) > sc^P_1(c_2) \ldots$ (for completeness, let unavailable candidates have score 0). Also, recall that $D^1_{ij} = sc^P_1(c_i) - sc^P_1(c_j)$. The following lemma[1] asserts that the top two scores are likely to be close to one another.

**Lemma 7.** $D^1_{12} = O(\sqrt{\frac{n \log^2 \log m}{m \log m}}) = o(\sqrt{\frac{n}{m}})$ *w.h.p.*

*Proof.* Let $A \subseteq C$ be the availability set ($|A| = \tilde{m}$), and partition $A$ into two (roughly) equal size sets: $A_1, A_2 \subset A$, such that $|A_1| = \lfloor \tilde{m}/2 \rfloor, |A_2| = \lceil \tilde{m}/2 \rceil$. Define two random variables: $t_1 = \max_{c \in A_1} sc^P_1(c), t_2 = \max_{c \in A_2} sc^P_1(c)$. It it easy to see that $D^1_{1,2} \leq |t_1 - t_2|$, so we prove the claim by upper-bounding the r.h.s. of the inequality. The number of the votes in $V_1$ that rank candidates in $A_1$ ($A_2$) first is bounded away from $n_1/2$ by $O(\sqrt{n_1})$ w.h.p. So the score of each candidate in $A_1$ and $A_2$ is distributed according to a typical balls-and-bins process, in which $n_1/2 \pm o(n_1)$ balls are thrown into $\tilde{m}/2$ bins, at random. Using Thm. 1 of [20], we have $|t_i - \mathbb{E}[t_i]| = \Theta\left(\sqrt{\frac{n_1 \log \tilde{m}}{\tilde{m}}} \sqrt{(1 - (1+\epsilon)\frac{\log \log \tilde{m}}{2 \log \tilde{m}})}\right)$, for $\epsilon > 0$ w.h.p., for $i = 1, 2$. Using our bounds on $\tilde{m}, n_1$, and the approximation $\sqrt{1-x} = 1 - \Theta(x)$, we derive $|t_1 - t_2| = O(\sqrt{\frac{n \cdot \log m}{m}} \frac{\log log m}{\log m}) = O(\frac{n \log^2 \log m}{\log m})$, w.h.p. □

**Lemma 8.** *Let $k = o(\log m)$. Then $D^2_{2,1} = \Omega(\sqrt{n/m})$ with constant probability.*

The proof is similar to that of Lemma 4. □

To summarize, we see that top-$k$ voting can be very effective for plurality voting with the possibility of unavailable candidates under the impartial culture model, requiring elicitation of only the $O(\log m)$ most-preferred candidates from voters to ensure the correct winner w.h.p. (this upper bound is tight). If one wants worst case assurances, this same bound suffices for "small" elections (with a number of voters polynomial in $n$); but for "large" elections (with an exponential number of voters), top-$k$ voting offers no savings.

## 4 Top-$k$ Voting and Borda Scoring

We now turn our attention to Borda scoring, and provide similar results. As with plurality, we begin with a worst-case lower bound on $k$. We note that the following result follows quite directly from a general result on the (deterministic) communication complexity of any rank-based voting rule, which Conitzer and Sandholm [7] show to require $O(nm \log m)$ bits in the worst case. However, we provide a direct construction for Borda.

**Theorem 9** (Worst case lower bound)**.** *Top-$k$ voting requires $k = \Omega(m)$ to determine the correct Borda winner w.h.p. in the worst-case, even when $p = 1$.*

*Proof.* Assume for simplicity that $|C|$ is odd and larger than 5. Let $C = \{c\} \cup A$ for some designated candidate $c$. Let $\pi$ be an arbitrary ordering of $A$, and $\pi^r$ its reverse. Let $(\pi_1, \pi_2)$ be a profile with two votes, s.t. $\pi_1$ and $\pi_2$ are obtained by placing

[1] We thank Neal Young for the idea of the proof.

$c$ between candidates ranked $(m-1)/2-1$ and $(m-1)/2$ in $\pi$ and $\pi^r$. If $k=(m-1)/2-1$, $c$ is not the Borda winner, though its average score is $(m+1)/2$, whereas the average score of all other candidates is $(m-1)/2$. □

We now provide an average-case lower bound on $k$ under the impartial culture assumption.

**Theorem 10** (Avg. case lower bound)**.** *If $n=\Omega(m^3\cdot\log m)$, then $k=\Omega(m/\log m)$ is necessary for top-$k$ voting to produce the true Borda winner w.h.p. under impartial culture, even when $p=1$.*

**Sketch of Proof** We provide a brief proof sketch. The proof idea is similar to that for plurality: we upper bound the *observed* difference in score between the winner and any other candidate under top-$k$ voting. We then show that with constant probability the score difference between winner and the second-highest candidate is eliminated as a result of discounted votes. Given the *true* Borda scores $sc_i^B(\cdot)$ of the candidates in vote $\pi_i$, let $\alpha_i(c)=sc_i^B(c)$ if $sc_i^B(c)\geq m-k$, and $\alpha_i=0$, otherwise. That is, $\alpha_i(c)$ is the Borda score of $c$ according to top-$k$ voting. Similarly, let $\beta_i(c)=sc_i^B(c)$ if $sc_i^B(c)<m-k$ and $\beta_i(c)=0$ otherwise; i.e., the extra score "lost" due to top-$k$ voting. We let $\alpha(c)=\sum_{i\in N}\alpha_i(c)$ and $\beta(c)=\sum_{i\in N}\beta_i(c)$. Finally, for two distinct candidates $c,c'\in C$, $D^T(c,c')=\alpha(c)-\alpha(c')$, and $D^B(c,c')=\beta(c)-\beta(c')$. We argue that if $c$ and $c'$ are the highest and second-highest scoring candidates under top-$k$ voting, if $k=o(m/\log m)$, $D^T(c,c')<D^B(c',c)$ with constant probability.

**Lemma 11.** *If $k=o(m/\log m)$, then for all $c,c'\in C$, $D^T(c,c')=o(\sqrt{n}\frac{m}{\log m})$ with high probability.*

The above lemma can be proved by bounding the variance of $D^T(c,c')$ and applying the Bernstein inequality.

Next, we claim that difference in uncounted scores due to truncation can be greater than this observed gap between the highest and second highest scores, impacting the true winner.

**Lemma 12.** *If $k=O(m/\log m)$ then $D^B(c',c)=\Omega(m\sqrt{n})$ with constant probability, where $c$ and $c'$ are the candidates with the highest and second highest scores.*

The proof of Lemma 12 is similar to Lemma 4 (albeit somewhat more involved) and requires the bounding of the second and third moment of $D^B(c',c)$ and making use of the Berry-Esseen theorem.

Combining Lemma 11 and Lemma 12 we get that $D(c,c')=D^B(c,c')+D^T(c,c')<0$ with constant probability, which proves the theorem. □

To summarize, top-$k$ voting cannot ease the elicitation burden in Borda elections in the worst case. Under impartial culture, there is hope for *some* elicitation savings for elections of reasonable size, as indicated by our lower bound of $k=\Omega(m/\log m)$, which suggests that $O(m/\log m)$ might suffice. But these savings are not nearly as substantial as in the case of plurality, nor are they guaranteed without a matching upper bound. A matching upper bound, or a stronger lower bound—for instance, perhaps our proof could be strengthened to give a lower bound of $\Omega(m)$—is an important result needed to complete the picture regarding Borda under impartial culture. Despite this, we will see below that top-$k$ voting can, in fact, help substantially in Borda voting under other, more realistic preference distributions.

## 5 Zero-elicitation Protocols

It is widely recognized that the impartial culture assumption does not provide a realistic model of real-world preferences or voting situations [21]. For this reason, exploring the ability to limit elicitation under other, more realistic probabilistic models of voter preference is of great import. We consider one such model in this work, namely the Mallows model, since it allows us to generalize the impartial culture model (which is a special case) by simply varying the dispersion or degree of concentration of voter preferences in a natural way. While we do not claim that the Mallows model is an ideal model for all social choice situations (though it serves as an important backbone for mixture models of preferences [19; 4; 14]), it serves as an important starting point for the broader investigation of top-$k$ voting.

In this section, we theoretically analyze the special case of *zero elicitation*, in other words, setting $k=0$ in top-$k$ voting, under Mallows model distributions. Specifically, we ask how concentrated voter preferences need to be—what dispersion values $\varphi$ suffice—to ensure that correct plurality and Borda winners can be selected w.h.p. *without eliciting any information from voters*. For ease of presentation, we assume $p=1$ (i.e., all candidates are available); however, our proofs can be modified to accommodate $p<1$, using simple applications of Chernoff and union bounds to account for missing candidates. In the next section, we empirically analyze top-$k$ voting for both zero elicitation and more general values of $k$ under Mallows models.

Assume a Mallows model $(\hat{\pi},\varphi)$ over $m$ candidates $C$. With no elicitation, the candidate with the expected highest (plurality or Borda) score is obviously the highest ranked candidate $\hat{\pi}(1)$, and it has the highest probability of winning (assuming $\varphi<1$, otherwise all candidates are equally likely to be winners). Under plurality voting, we can show that with a large enough voter population, this approach performs well.

**Theorem 13.** *If $n=\Omega\left(\frac{\log m(1-\varphi^m)}{(1-\varphi)^3}\right)$, then the highest ranked candidate $\hat{\pi}(1)$ is the plurality winner w.h.p.*

Thm. 13 can be proved using the Bernstein inequality and union bound to bound the probability that the highest-ranked candidate in $\hat{\pi}$ is dominated by another.

We can derive a similar bound for Borda voting.

**Theorem 14.** *If $n\geq\Gamma(\varphi)\ln m$, where $\Gamma(\varphi)=(8(1+\varphi)^2(1-\varphi)^3+(1+\varphi))/(1-\varphi)^7$, then the highest ranked candidate $\hat{\pi}(1)$ is the Borda winner w.h.p.*

As with Thm. 13, Thm. 14 makes use of the Bernstein theorem, more precisely, a version that requires a refined analysis of the distribution of Borda scores, and involves bounding the $k$'th moment of the difference in the Borda scores of the first and $j$'th candidates.

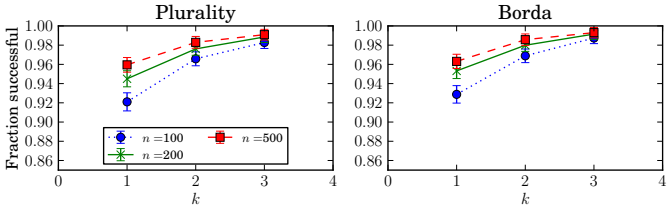


**Figure 1**: Correctness of top-$k$ voting: $m = 10$, varying $k$ and $n$.

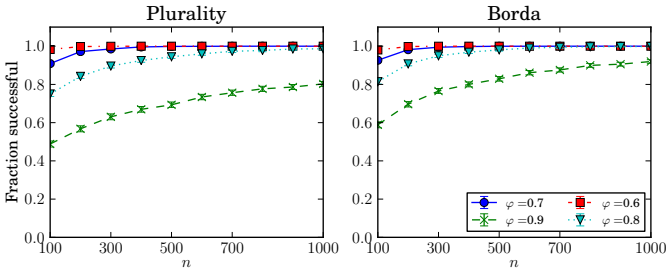


**Figure 2**: Correctness of zero elicitation: $m = 10$, varying $n, \varphi$.

## 6 Empirical Results

The bounds above provide some theoretical justification for the use of top-$k$ voting; however, they do not prescribe precise values for the choice of $k$ with respect to specific priors and election sizes $(m, n)$. In this section we present simulation results for small elections with $m = 10$ candidates and $n = 100$ to 5000 voters to illustrate the probability of correct winner selection in both plurality and Borda elections using top-$k$ voting for several values of $k$ (including zero elicitation), under Mallows models with a range of dispersions $\varphi$. In our experiments, we generate 10,000 random preference profiles for each parameter setting by drawing voter rankings i.i.d. from the appropriate Mallows model, and measure the fraction of such profiles in which top-$k$ gives the true winning candidate. (We assume $p = 0.5$ throughout, except for the results concerning zero elicitation.)

For top-$k$ experiments dispersion $\varphi < 0.7$, winner prediction was essentially perfect, even with $k = 1$, regardless of the other parameters. Fig. 1 shows the success rate of top-$k$ voting for plurality and Borda for $\varphi = 0.7$ and $m = 10$ as we vary $k$ and the number of voters. In all cases top-$k$ converges to the correct prediction, and is near perfect when $k = 3$. With more voters the performance is better, but the dependence is slight and almost negligible for $k = 3$.

For zero elicitation, we measured how often the first-ranked candidate in the Mallows reference ranking is the true election winner under plurality and Borda. We set $m = 10$, $p = 1$ for simplicity, vary $\varphi$ and $n$, and show results averaged over 10,000 elections each in Fig. 2. For $\varphi \leq 0.8$, predictions are near-perfect for $n \geq 700$, and $\varphi \leq 0.7$, $n \geq 400$ suffices for near 100% accuracy. We note that results are better for Borda than for plurality. For populations with an extremely high degree of dispersion ($\varphi = 0.9$), plurality success rate is only 0.8 at $n = 1000$, and Borda success rate is only 0.92. This conforms to our theoretical bounds in the sense that the success probability depends exponentially on $\varphi$, which means that it decreases dramatically for larger values of $\varphi$.

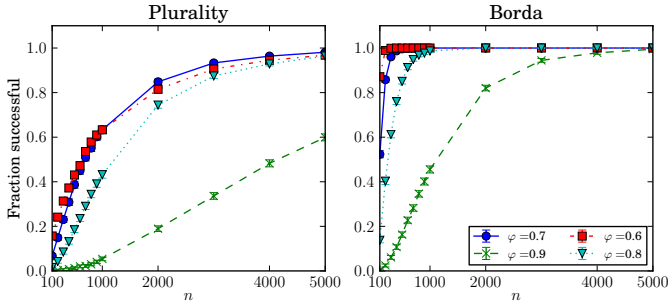


**Figure 3**: Reconstruction rates of zero-elicitation: $m = 10$, varying $n, \varphi$.

For zero elicitation, we also measured how often the voting method, using scores to reconstruct a societal ranking, accurately predicted the entire reference ranking $\hat{\pi}$ from the Mallows model. Results are depicted in Fig. 3. Unsurprisingly, the probability of complete ranking reconstruction is significantly lower than the probability to correctly forecast the winner. However, $n = 1000$ allows almost perfect reconstruction under Borda for $\varphi \leq 0.8$. Notice that the difference between plurality and Borda is even more pronounced than in winner prediction. Under Borda, $n = 5000$ suffices to reconstruct the entire ranking even for $\varphi = 0.9$, while for plurality, results for $\varphi = 0.9$ are much worse (about 0.6), and even for $\varphi = 0.7$ do not reach 100%.

## 7 Conclusions

We have provided a detailed analysis of top-$k$ voting, allowing for the possibility of unavailable candidates, for both plurality and Borda voting. Our theoretical results place bounds (in some cases tight) on the required values of $k$ needed to determine the correct winner w.h.p., in both a worst-case sense and an average-case sense under impartial culture, and also showed under what conditions zero elicitation admits correct winner prediction under Mallows models. Our empirical results further demonstrated that relatively small values of $k$ work well in practice. Even zero elicitation shows strong promise when preferences exhibit just mild degrees of correlation for elections with a sufficient number of voters.

There are a number of interesting directions for future research. Extending our analysis to other voting rules is of great interest. For example, preliminary results suggest that Copeland exhibits behavior similar to Borda, requiring large $k$ for impartial culture; do certain voting rules exhibit behavior that is intermediate between plurality and Borda? Extending our analysis to a richer class of realistic preference distributions, such as the Plackett-Luce model, or Mallows mixtures, is an important next step, as is testing our approach on real data sets.

A third direction is the investigation of multi-round elicitation protocols [16]. In such protocols, voting data is elicited in stages, and the protocol terminates when the winner can be determined with high probability. Such protocols are adaptive and dynamic, eliciting information in a given stage conditioned on information gleaned in earlier stages. An important question is whether it is possible to elicit less information on average with such a protocol.

## References


[1] Katherine A. Baldiga and Jerry R. Green. Assent-maximizing social choice. *Social Choice and Welfare*, pages 1–22, 2011. Online First.

[2] John Bartholdi III, Craig Tovey, and Michael Trick. How hard is it to control an election? *Social Choice and Welfare*, 16(8-9):27–40, 1992.

[3] Craig Boutilier, Jérôme Lang, Joel Oren, and Héctor Palacios. Robust winners and winner determination policies under candidate uncertainty. In *Proceedings of the Fourth International Workshop on Computational Social Choice (COMSOC-2012)*, Kraków, Poland, 2012.

[4] Ludwig M. Busse, Peter Orbanz, and Joachim M. Buhmann. Cluster analysis of heterogeneous rank data. In *Proceedings of the Twenty-fourth International Conference on Machine Learning (ICML-07)*, pages 113–120, 2007.

[5] Yann Chevaleyre, Jérôme Lang, Nicolas Maudet, and Jérôme Monnot. Possible winners when new candidates are added: The case of scoring rules. In *Proceedings of the Twenty-fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, pages 762–767, Atlanta, GA, 2010.

[6] Vincent Conitzer. Eliciting single-peaked preferences using comparison queries. *Journal of Artificial Intelligence Research*, 35:161–191, 2009.

[7] Vincent Conitzer and Tuomas Sandholm. Communication complexity of common voting rules. In *Proceedings of the Sixth ACM Conference on Electronic Commerce (EC'05)*, pages 78–87, Vancouver, 2005.

[8] Ning Ding and Fangzhen Lin. Voting with partial information: Minimal sets of questions to decide an outcome. In *Proceedings of the Fourth International Workshop on Computational Social Choice (COMSOC-2012)*, Kraków, Poland, 2012.

[9] Edith Hemaspaandra, Lane Hemaspaandra, and Jörg Rothe. Anyone but him: The complexity of precluding an alternative. *Artificial Intelligence*, 171(5-6):255–285, 2007.

[10] Meir Kalech, Sarit Kraus, Gal A. Kaminka, and Claudia V. Goldman. Practical voting rules with partial information. *Journal of Autonomous Agents and Multi-Agent Systems*, 22(1):151–182, 2011.

[11] Kathrin Konczak and Jérôme Lang. Voting procedures with incomplete preferences. In *IJCAI-05 Workshop on Advances in Preference Handling*, pages 124–129, Edinburgh, 2005.

[12] V. Yu. Korolev and I. G. Shevtsova. On the upper bound for the absolute constant in the Berry-Esseen inequality. *Theory Probab. Appl.*, 54(4):628–658, 2010.

[13] Tyler Lu and Craig Boutilier. The unavailable candidate model: A decision-theoretic view of social choice. In *Proceedings of the Eleventh ACM Conference on Electronic Commerce (EC'10)*, pages 263–274, Cambridge, MA, 2010.

[14] Tyler Lu and Craig Boutilier. Learning Mallows models with pairwise preferences. In *Proceedings of the Twenty-eighth International Conference on Machine Learning (ICML-11)*, pages 145–152, Bellevue, WA, 2011.

[15] Tyler Lu and Craig Boutilier. Robust approximation and incremental elicitation in voting protocols. In *Proceedings of the Twenty-second International Joint Conference on Artificial Intelligence (IJCAI-11)*, pages 287–293, Barcelona, 2011.

[16] Tyler Lu and Craig Boutilier. Vote elicitation with probabilistic preference models: Empirical estimation and cost tradeoffs. In *Proceedings of the Second International Conference on Algorithmic Decision Theory (ADT-11)*, pages 135–149, Piscataway, NJ, 2011.

[17] Colin L. Mallows. Non-null ranking models. *Biometrika*, 44:114–130, 1957.

[18] John I. Marden. *Analyzing and modeling rank data*. Chapman and Hall, London, 1995.

[19] Thomas Brendan Murphy and Donal Martin. Mixtures of distance-based models for ranking data. *Computational Statistics and Data Analysis*, 41:645–655, January 2003.

[20] Martin Raab and Angelika Steger. "balls into bins" - a simple and tight analysis. In *Proceedings of the Second International Workshop on Randomization and Approximation Techniques in Computer Science*, RANDOM '98, pages 159–170, London, UK, UK, 1998. Springer-Verlag.

[21] Michel Regenwetter, Bernard Grofman, A. A. J. Marley, and Ilia Tsetlin. *Behavioral Social Choice: Probabilistic Models, Statistical Inference, and Applications*. Cambridge University Press, Cambridge, 2006.

[22] Lirong Xia and Vincent Conitzer. Determining possible and necessary winners under common voting rules given partial orders. In *Proceedings of the Twenty-third AAAI Conference on Artificial Intelligence (AAAI-08)*, pages 202–207, Chicago, 2008.