PROVIDENCE: a Flexible Round-by-Round Risk-Limiting Audit

Your N. Here

Your Institution

Second Name

Second Institution

Abstract

A Risk-Limiting Audit (RLA) is a statistical election tabulation audit with a rigorous error guarantee. Real ballot polling RLAs draw ballots in rounds of multiple ballots each. For practical round sizes, recently-proposed RLA MINERVA draws fewer ballots than commonly-used RLA BRAVO (half as many for large round sizes) but is more rigid because it requires that the round schedule not change once the audit begins.

We present PROVIDENCE, an audit with the efficiency of MINERVA and flexibility of BRAVO. We prove that PROVIDENCE is risk-limiting in the presence of an adversary who can choose subsequent round sizes given knowledge of previous samples. We describe a measure of audit workload—as a function of the number of rounds, precincts touched and ballots drawn—and quantify the problem of obtaining a misleading audit outcome when rounds are too small. We present simulation results demonstrating the superiority of PROVIDENCE using these measures.

We describe the use of PROVIDENCE by the Rhode Island Board of Elections in a tabulation audit of the 2021 election. Our implementation of PROVIDENCE in the open source R2B2 library should be useful to the states of Virginia, Georgia and Pennsylvania, which plan pre-certification RLAs for the 2022 general election.

1 Introduction

The literature contains numerous descriptions of vulnerabilities in deployed voting systems, and it is not possible to be certain that any system, however well-designed, will perform as expected in all instances. For this reason, *evidence-based elections* [13] aim to produce trustworthy and compelling evidence of the correctness of election outcomes, enabling the detection of problems with high probability. One way to implement an evidence-based election is to use a well-curated voter-verified paper trail, compliance audits, and a rigorous tabulation audit of the election outcome, known as a risk-limiting audit (RLA) [6]. An RLA is an audit which guarantees that the probability of concluding that an election outcome is correct, given that it is not, is below a pre-determined value known as the risk limit of the audit, independent of the true, unknown vote distribution of the underlying election. Over a dozen states in the US have seriously explored the use of RLAs—some have pilot programs, some allow RLAs to satisfy a general audit requirement and some have RLAs in statute.

This paper concerns ballot-polling RLAs, which require a large number of ballots relative to comparison RLAs but do not rely on any special features of the election technology. Since comparison RLAs are not always feasible, ballot-polling audits remain an important resource and have been used in a number of US state pilots (California, Georgia, Indiana, Michigan, Ohio, Pennsylvania, Virginia and elsewhere). In the general ballot-polling RLA, a number of ballots are drawn and tallied in what is termed a *round* of ballots [17]. A statistical measure is then computed to determine whether there is sufficient evidence to declare the election outcome correct within the pre-determined risk limit. If so, the audit stops; else another round is drawn. Election officials would typically decide to do a full manual hand count if the audit does not stop in spite of drawing a large number of ballots. Because the decision to stop or draw more ballots is made after drawing a round of ballots, the audit is termed a *round-by-round (R2)* audit. The special case when round size is one—that is, stopping decisions are made after each ballot draw—is a *ballot-by-ballot (B2)* audit.

The BRAVO audit is designed for use as a B2 audit: it requires the smallest expected number of ballots when the true tally of the underlying election is as announced and stopping decisions are made after each ballot draw. In practice, election officials draw many ballots at once, and the BRAVO stopping rule needs to be modified for use in an R2 audit that is not B2. There are two obvious approaches. The B2 stopping condition can be applied once at the end of each round: End-of Round (EoR) BRAVO. Alternatively, the order of ballots in the sample can be tracked by election officials and the

B2 BRAVO stopping condition can be applied retroactively after each ballot drawn: Selection-Ordered (SO) BRAVO. SO BRAVO requires fewer ballots on average than EoR BRAVO but requires the work of tracking the order of ballots rather than just their tally.

MINERVA was designed for R2 audits and applies its stopping rule once for each round. Thus it does not require the tracking of ballots that SO BRAVO does. Zagórski *et al.* [17] prove that MINERVA is a risk-limiting audit and requires fewer ballots to be sampled than EoR BRAVO when an audit is performed in rounds, the two audits have the same pre-determined (before any ballots are drawn) round schedule and the underlying election is as announced. They also present first-round simulations which show that MINERVA draws fewer ballots than SO BRAVO in the first round for first round sizes with a large probability of stopping when the (true) underlying election is as announced. Broadrick *et al.* provide further simulations that show MINERVA requires fewer ballots over multiple rounds and for lower stopping probability.

While more efficient than BRAVO, MINERVA requires that the round schedule is fixed in advance of the audit. This may lead to significant unnecessary work. For example, suppose the fixed round schedule for MINERVA is 10,000 ballots per round (a reasonable number in a state-wide contest with a narrow margin), and the first round sample finds a number of ballots for the winner just short of that necessary for the audit to stop. It then may be sufficient to draw a very small second round and still stop with high probability in the second round, but the additional 10,000 ballots must be drawn. In contrast, subsequent BRAVO round sizes can be chosen based on preceding samples.

An open question is whether a ballot polling RLA exists with the efficiency of MINERVA and this flexibility of BRAVO.

1.1 Our Contributions

We present PROVIDENCE, and provide the following:

1. Proof that PROVIDENCE is an RLA and resistant to an adversary who can choose subsequent round sizes with knowledge of previous samples

2. Simulations of PROVIDENCE, MINERVA, SO BRAVO, and EoR BRAVO which show that PROVIDENCE has similar efficiency to MINERVA, both greater than either implementation of BRAVO

3. Results and analysis from the use of PROVIDENCE in a pilot audit in Rhode Island

4. Open source implementation of PROVIDENCE

2 Related work

The BRAVO audit [7] is a well-known ballot polling audit which has been used in numerous pilot and real audits. When used to audit a two-candidate election, it is an instance of Wald's sequential probability ratio test (SPRT) [15], and inherits the SPRT property of being the most efficient test (requiring the smallest expected number of ballots) if the election is as announced. The model for BRAVO and the SPRT is, however, that of a sequential audit: a sample of size one is drawn, and a decision of whether to stop the audit or not is taken. Real election audits invest in drawing large numbers of ballot, called rounds, before making stopping decisions because sequentially sampling individual ballots has significant overhead (unsealing storage boxes and searching for individual ballots). It is possible to apply BRAVO to the sequence of ballots in a round if the sequential order is retained. This is not, however, the most efficient possible use of the drawn sample because information in consequent ballots is ignored when applying BRAVO to ballots that were drawn earlier in the sample.

We do know a great deal about the properties of BRAVO. The risk limiting property of BRAVO follows from the similar property of the SPRT. Stopping probabilities for BRAVO may be estimated as implemented in [14]; this method is due to Mark Lindeman and uses quadratic approximations. A later method for stopping probability estimates presented by Zagórski *et al.* [16, 17] uses a similar technique for narrow margins and a separate algorithm for wider margins, the results of which match simulation results reported by Lindeman *et al.* [7, Table 1].

The MINERVA audit [16, 17] was developed for large first round sizes which enable election officials to be done in one round with large probability. It uses information from the entire sample, and has been proven to be risk limiting when the round schedule for the audit is determined before the audit begins. That is, information about the actual ballots drawn in the first round cannot inform future round sizes. First-round sizes for a 0.9 stopping probability when the election is as announced have been computed for a wide range of margins and are smaller than those for EoR and SO BRAVO. First round simulations of MINERVA [16] demonstrate that its first-round properties—regarding the probabilities of stopping when the underlying election is tied and when it is as announced—are as predicted for first round sizes with stopping probability 0.9. Additional simulations [3] have shown that MINERVA requires fewer ballots than EoR and SO BRAVO over multiple rounds and for smaller stopping probability. As expected, the advantage of MINERVA decreases for smaller stopping probability (smaller round sizes) as such round schedules approach the B2 round schedule (1,1,1,...) for which BRAVO is known to be most efficient.

Ballot polling audit simulations have been used to familiarize election officials and the public with the approach [12] and

to investigate the efficiency of audits [2,5,8,9]. Some workload measurements have been made [4]. While total ballots sampled can give naive workload estimates [11], Bernhard presents a more complex workload estimation model [1].

2.1 Model

An audit $\mathcal{A}$ is a function that takes as input the sample of ballots and outputs either (1) *Correct: stop the audit* or (2) *Undetermined: sample more ballots*. BRAVO and MINERVA are modeled as binary hypothesis tests where the null hypothesis H_0 corresponds to a true underlying tie and the alternative hypothesis H_a corresponds to the underyling election being as announced. (With an even total number of ballots, H_0 corresponds to the announced loser winning by one ballot.) Thus the null hypothesis is the incorrectly announced outcome which is most difficult to detect. A risk-limiting audit requires that in the event of an underlying tie, the probability of failing to detect the error is below the risk limit.

Definition 1 (Risk Limiting Audit (α-RLA))**.** *An audit $\mathcal{A}$ is a Risk Limiting Audit with risk limit α iff for sample X*

$$Pr[\mathcal{A}(X) = Correct \,|\, H_0] \leq \alpha$$

The stopping conditions of BRAVO and MINERVA rely on the following ratios.

Definition 2 (BRAVO Ratio)**.** *The* BRAVO *audit uses the ratio σ. Consider a sample size of n ballots with k for the reported winner. The proportion of ballots for the reported winner under the alternative hypothesis and null hypothesis are p_a and p_0 respectively.*

$$\sigma(k, p_a, p_0, n) \triangleq \frac{p_a^k (1-p_a)^{n-k}}{p_0^k (1-p_0)^{n-k}} \tag{1}$$

In BRAVO, $p_0 = \frac{1}{2}$. A BRAVO audit outputs correct if and only if

$$\sigma(k, p_a, \frac{1}{2}, n) \geq \frac{1}{\alpha}.$$

If testing the BRAVO stopping condition after each individual ballot is drawn (a B2 BRAVO audit), σ is equivalent to the likelihood ratio:

$$\frac{Pr[K=k|H_a,n]}{Pr[K=k|H_0,n]} = \frac{\binom{n}{k} p_a^k (1-p_a)^{n-k}}{\binom{n}{k}(\frac{1}{2})^n} = \sigma(k, p_a, \frac{1}{2}, n)$$

It now becomes useful to have shorthand for a sequence of round sizes and a sequence of winner ballot tallies. We use:

$$\boldsymbol{k_j} \triangleq (k_1, k_2, \ldots, k_j)$$

$$\boldsymbol{n_j} \triangleq (n_1, n_2, \ldots, n_j)$$

Where BRAVO uses a ratio of points on conditional probability distributions, MINERVA uses a ratio of *tails* of conditional probability distributions.

Definition 3 (MINERVA Ratio)**.** *The R2* MINERVA *audit uses the ratio τ_j. We use cumulative round sizes $\boldsymbol{n_j}$, with corresponding $\boldsymbol{k_j}$ ballots for the reported winner in reach round. The proportion of ballots for the reported winner under the alternative hypothesis and null hypothesis are p_a and p_0 respectively.*

$$\tau_j(k_j, p_a, p_0, \boldsymbol{n_j}, \alpha) \triangleq \frac{Pr[K_j \geq k_j \wedge \forall_{i<j}(\mathcal{A}(X_i) \neq Correct) \,|\, H_a, \boldsymbol{n_j}]}{Pr[K_j \geq k_j \wedge \forall_{i<j}(\mathcal{A}(X_i) \neq Correct) \,|\, H_0, \boldsymbol{n_j}]} \tag{2}$$

3 PROVIDENCE

We now introduce the stopping condition of PROVIDENCE. (TODO: add some more motivation and description of Providence... where minerva insists on a bound on the average risk providence only bounds the current path of the audit... it is nice because it sort of combines the main ratios of bravo and minerva)

Definition 4 ($(\alpha, p_a, p_0, k_{j-1}, n_{j-1}, n_j)$-PROVIDENCE)**.** *For cumulative round size n_i for round i and a cumulative k_i ballots for the reported winner found in round i, the R2* PROVIDENCE *stopping rule for the j^{th} round is:*

$$\mathcal{A}(X_j) = \begin{cases} Correct & \omega_j(k_j, k_{j-1}, p_a, p_0, n_j, n_{j-1}) \geq \frac{1}{\alpha} \\ Undetermined & else \end{cases}$$

where $\omega_1 \triangleq \tau_1$ and for $j \geq 2$, we define ω_j as follows:

$$\omega_j(k_j, k_{j-1}, p_a, p_0, n_j, n_{j-1}) \triangleq \sigma(k_{j-1}, p_a, p_0, n_{j-1}) \cdot \tau_1(k_j - k_{j-1}, p_a, p_0, n_j - n_{j-1}) \tag{3}$$

Notice that for $j \geq 2$, unlike τ_j, computing ω_j requires no convolution. We now we prove that PROVIDENCE is risk-limiting using lemmas from basic algebra in Appendix A.

Theorem 1. *An $(\alpha, p_a, p_0, k_{j-1}, n_{j-1}, n_j)$-*PROVIDENCE *audit is an α-RLA.*

Proof. Let $\mathcal{A} = (\alpha, p_a, p_0, k_{j-1}, n_{j-1}, n_j)$-PROVIDENCE. Let $\boldsymbol{n_j}$ be the cumulative roundsizes used in this audit, with corresponding cumulative tallies of ballots for the reported winner $\boldsymbol{k_j}$. For round $j = 1$, by Definitions 4 and 3, we see that the $\mathcal{A} =$ Correct (the audit stops) only when

$$\tau_1(k_1, p_a, p_0, n_1) = \frac{Pr[K_1 \geq k_1 \,|\, H_a, n_1]}{Pr[K_1 \geq k_1 \,|\, H_0, n_1]} \geq \frac{1}{\alpha}.$$

By Lemma 5, we see that this is equivalent to the following:

$$\frac{Pr[K_1 \geq k_{min,1} \,|\, H_a, n_1]}{Pr[K_1 \geq k_{min,1} \,|\, H_0, n_1]} \geq \frac{1}{\alpha}.$$

For any round $j \geq 2$, by Definition 4 and Lemma 5, $\mathcal{A} =$ Correct (the audit stops) only when

$$\omega_j(k_j, k_{j-1}, p_a, p_0, n_j, n_{j-1}, \alpha) \triangleq \sigma(k_{j-1}, p_a, p_0, n_{j-1}) \cdot \tau_1(k_j - k_{j-1}, p_a, p_0,$$

By Lemma 6 and Definition 3, this is equivalent to

$$\frac{\Pr[\boldsymbol{k}_{j-1} \mid H_a] \cdot Pr[K_j \geq k_j \mid \boldsymbol{k}_{j-1}, H_a, nj]}{\Pr[\boldsymbol{k}_{j-1} \mid H_0] \cdot Pr[K_j \geq k_j \mid \boldsymbol{k}_{j-1}, H_0, nj]} \geq \frac{1}{\alpha}.$$

By Lemma 5 and Definition 4, we see that there exists a $k_{min,j} \leq k_j$ for which

$$\frac{\Pr[\boldsymbol{k}_{j-1} \mid H_a] \cdot Pr[K_j \geq k_j \mid \boldsymbol{k}_{j-1}, H_a, nj]}{\Pr[\boldsymbol{k}_{j-1} \mid H_0] \cdot Pr[K_j \geq k_j \mid \boldsymbol{k}_{j-1}, H_0, nj]} \geq$$

$$\frac{\Pr[\boldsymbol{k}_{j-1} \mid H_a] \cdot Pr[K_j \geq k_{min,j} \mid \boldsymbol{k}_{j-1}, H_a, nj]}{\Pr[\boldsymbol{k}_{j-1} \mid H_0] \cdot Pr[K_j \geq k_{min,j} \mid \boldsymbol{k}_{j-1}, H_0, nj]} \geq$$

$$\frac{\Pr[\boldsymbol{k}_{j-1} \mid H_a] \cdot Pr[K_j \geq k_j \mid \boldsymbol{k}_{j-1}, H_a, nj]}{\Pr[\boldsymbol{k}_{j-1} \mid H_0] \cdot Pr[K_j \geq k_j \mid \boldsymbol{k}_{j-1}, H_0, nj]} \geq \frac{1}{\alpha}.$$

Taking the sum over all possible audit histories, we get

$$\frac{\sum_{\boldsymbol{k}_j} \Pr[\boldsymbol{k}_{j-1} \mid H_a] \cdot Pr[K_j \geq k_{min,j} \mid \boldsymbol{k}_{j-1}, H_a, nj]}{\sum_{\boldsymbol{k}_j} \Pr[\boldsymbol{k}_{j-1} \mid H_0] \cdot Pr[K_j \geq k_{min,j} \mid \boldsymbol{k}_{j-1}, H_0, nj]} \geq \frac{1}{\alpha}.$$

Finally, because the total probability of stopping the audit under the alternative hypothesis is less than 1, we get

$$\frac{Pr[\mathcal{A} = \text{Correct} \mid H_a]}{Pr[\mathcal{A} = \text{Correct} \mid H_0]} \geq \frac{1}{\alpha}$$

$$Pr[\mathcal{A} = \text{Correct} \mid H_0] \leq Pr[\mathcal{A} = \text{Correct} \mid H_a] \cdot \alpha \leq \alpha.$$

□

3.1 Resistance against an adversary choosing round sizes

I don't feel very comfortable with how to word/present this argument... I have not yet made an attempt at doing so.

4 Simulations

We use simulations to provide additional evidence for theoretical claims and gain insight into audit behavior. As in [**?**], we use margins from the 2020 US Presidential election, statewide pairwise margins between the leading two candidates of 5% or more. Narrower margins are computationally expensive, especially for the simulations with an underyling tie which quickly increase in sample size. We use the simulator in the R2B2 software library [10]. We perform $10000 = 10^4$ trials per margin for both an underyling outcomed as announced and an underyling tie.

In the simulations with an underlying tie, the percent of audits that stop, as shown in Figure 1, give an estimate of maximum risk. For all margins, this estimated risk is less than the risk limit, supporting the claim that PROVIDENCE is risk-limiting.

Simulations with the underlying ballot distribution as announced provide insight into stopping probability and number

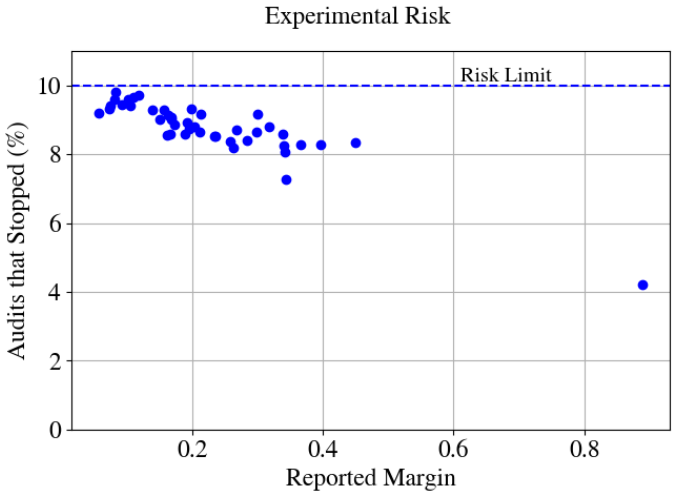

Figure 1: The percent of simulated audits PROVIDENCE with an underlying tie that stopped in any of the five simulated rounds. This percent gives an estimate of the risk of the PROVIDENCE.

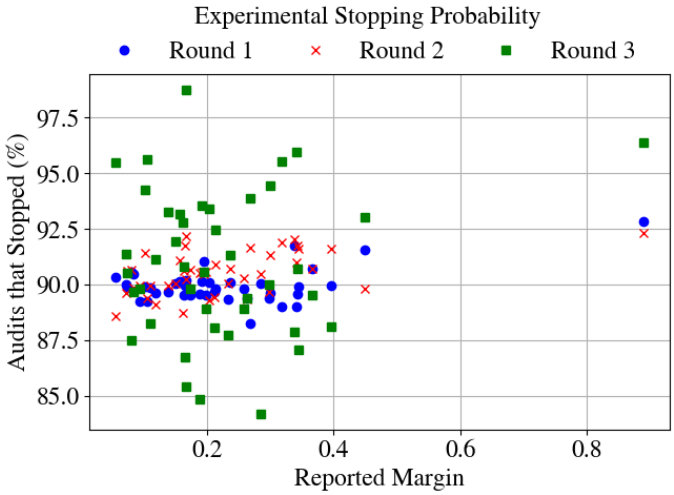

Figure 2: For the simulated PROVIDENCE audits with underlying margin as announced, the percent of audits that stopped among those that entered a round. This percent gives an estimate of the stopping probability conditioned on the sample of the previous round. The average percent for rounds 1, 2, and 3 is 89.96%, 90.52%, and 90.98% respectively. We show only the first three rounds since so few audits make it to rounds 4 and 5.

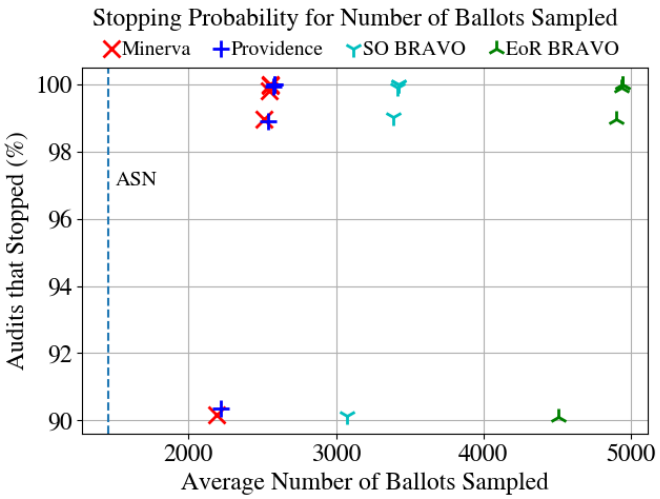

Figure 3: For all five rounds, the estimated stopping probability for average number of ballots drawn for PROVIDENCE, MINERVA, EoR BRAVO, and SO BRAVO.

of ballots drawn. Figure 2 shows that the stoping probabilities over the first rounds are near and slightly above 90% as expected since our software chose round sizes to give at least a 90% conditional stopping probability. Figure 3 shows that the probability of stopping as a function of number of ballots sampled. Points above (higher probability of stopping) and to the left (fewer ballots) represent more efficient audits. As shown, PROVIDENCE has comparable efficiency to MINERVA, while both are significantly more efficient than either implementation of BRAVO. Note that a difference in twice as many ballots could be negligible in a contest with a wide margin. In a contest with a narrow margin (in the 2020 US Presidential election, eight states had margins less than 3%) the difference in number of ballots sampled could be weeks of work. Section 6 discusses workload in more depth.

5 Pilot use

The Rhode Island Board of Elections performed a pilot audit in Providence in February 2022. The contest audited was a single yes-or-no question from the November 2021 special election. The question had announced margin 25.67%. The audit used risk-limit 10%. A first round size of 140 ballots with large probability of stopping (95%) was selected in order to give the potential for more interesting analysis afterwards; the large margin made a first round with large stopping probability practical. Selection order was tracked for the sake of analysis. As expected, the audit concluded in the first round with a PROVIDENCE risk of 4.18%. Table 1 shows risk measures for the drawn sample using MINERVA and BRAVO (both EoR and SO).

TODO: Add examples of how the audits perform for various hypothetical round schedules. I wait to do this until I'm done with the workload estimates since the examples here should be chosen to motivate that section.

ballots	PROVIDENCE	MINERVA	SO BRAVO	EoR BRAVO
140	**4.18%**	**4.18%**	**5.41%**	36.6%

Table 1: Risk measures for the drawn first round of 140 ballots in the Providence, RI pilot audit. Risks in bold meet the risk-limit (10%) and thus correspond to audits that would stop.

6 Audit workload

Some election audits have benefited from a one-and-done approach: draw a large sample with high probability of stopping in the first round and usually avoid a second round altogether. This is appealing for two reasons. Firstly, rounds have some overhead in both time and effort. Thus the time and person-hours of an audit grows not just with the number of ballots sampled but also with the number of rounds. Secondly, smaller first round sizes give a higher probability that the result after the first round is misleading in the sense that the true winner recieves has fewer votest than some other candidate according to the tally of the sampled ballots. On the other hand, a one-and-done audit may draw more ballots than are necessary; a more efficient round schedule could require less effort and time pre-certification. To evaluate the quality of various round schedules, we construct a simple workload model. Under this model we show how optimal round schedules can be chosen. We provide software that can be used by election officials to choose round schedules based on estimates of the model parameters like desired probability of a nonmisleading result.

As an example, we consider the US Presidential contest in the 2016 Virginia statewide general election. This contest had a margin of 5.3% between the two candidates with the most votes. Analytical approximation of the expected audit behavior (E_b and E_r) is challenging because the number of possible sequences of samples grows exponentially with the number of rounds. Therefore we use the typical approach of simulations, again with risk-limit 10%. We consider a simple round schedule, in which each round is selected to give the same probability of stopping, p. That is, if the audit does not stop in the first round, we select a second round size which, given the sample drawn in the first round, will again give a probability of stopping p in the second round. For this round schedule scheme, a one-and-done audit is achieved by choosing large p, say $p = .9$ or $p = .95$.[1] We run 10^4 trials for

[1] For this particular round schedule scheme, computing the expected number of rounds is possible analytically, but the expected number of ballots is

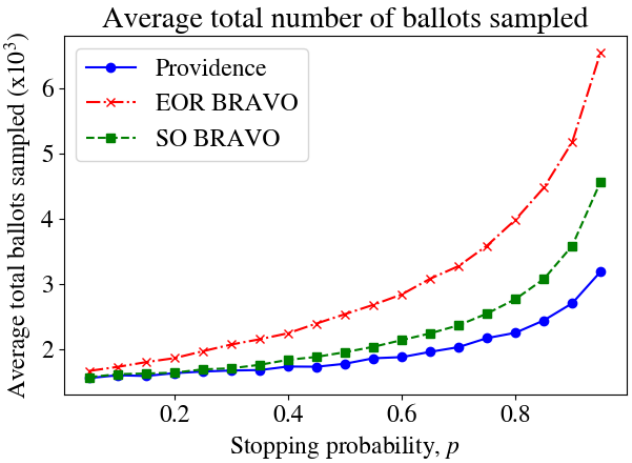

Figure 4: The total number of ballots sampled on average in our simulations for various round schedules parameterized by p the conditional stopping probability used to select each round size.

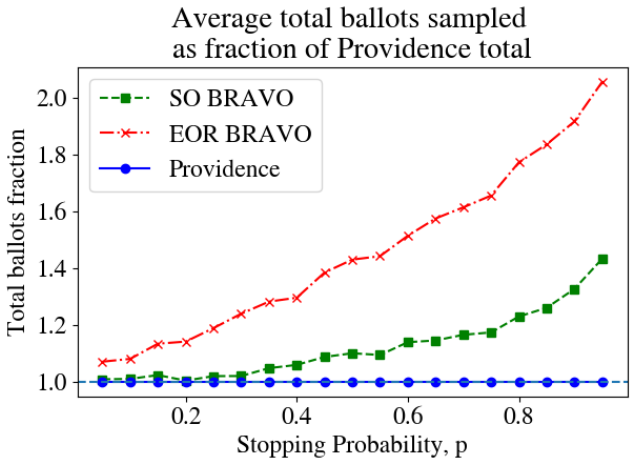

Figure 5: The total number of ballots sampled on average in our simulations given as a fraction of those sampled by PROVIDENCE, for various round schedules parameterized by p the conditional stopping probability used to select each round size.

each value of p, assuming the announced results are correct.

6.1 Person-hours

Average total ballots. The simplest workload models are a function of just the total number of a ballots sampled. Figure 4 shows the average total number of ballots sampled in our simulations for each value of p, which gives an estimate of the expected total number of ballots. Figure 5 gives the same number as a ratio of the PROVIDENCE values. It is straightforward to show that PROVIDENCE and both forms of BRAVO collapse to the same test in the case where each round is a single ballot. Figures 4 and 5 show that for larger stopping probabilities p (i.e. larger rounds), PROVIDENCE requires fewer ballots on average. In particular, the savings of PROVIDENCE become larger as p increases; for $p = .95$, EoR BRAVO and SO BRAVO require more than 2 and 1.4 times as many ballots as PROVIDENCE respectively.

Round overhead. It is clear that average number of ballots alone is an inadequate workload measure. (Consider a state conducting its audit by selecting a single ballot at random, notifying just the county where the ballot is located, and then waiting to hear back for the manual interpretation of the ballot before moving on to the next one. This of course is inefficient and is why audits are actually performed in rounds.)

In a US state-wide RLA, the state organizes the audit by determining the random sample and communicating with the counties, but election officials at the county level physically sample and inspect the ballots from their precincts. Therefore each audit round requires some number of person-hours for set up and communication between state and county. This overhead for a round includes choosing the round size, generating the random sample, and communicating that random sample to the counties, as well as the communication of the results back to the state afterwards.

Consequently, we now consider a model with a constant per-ballot cost c_b and a constant per-round cost c_r. So for an audit $\mathcal{A}$ with expected number of ballots E_b and expected number of rounds E_r, we estimate that the cost C of the audit is

$$C(\mathcal{A}) = E_b c_b + E_r c_r \tag{4}$$

For simplicity, (and without loss of generality), we assume the per ballot cost is one, $c_b = 1$. Then the per round cost c_r tells us the cost of a round as a number of ballots. We begin with $c_r = 1000$ as a conservative example. That is, we set the overhead of a round equal to the workload of sampling 1000 ballots. Based on available data [4], the time retrieving and analyzing each individual ballot is on the order of 75 seconds which means that $c_r = 1000$ is equivalent to roughly 20 person-hours of workload. This corresponds to about 15 minutes being spent per round in each of the 133 counties of Virginia, a clearly conservative workload estimate. As shown in Figure 6, lower average costs are achieved by selecting higher stopping probability; PROVIDENCE achieves the lowest minimum average cost is achieved at roughly 0.7.

Importantly, this gives us a way to estimate the expected cost, as well as which round schedule value p achieves it, for arbitrary round cost. For each round cost c_r, we produce a dataset analagous to that of Figure 6 and then find the minimum average cost achieved for each of the audits and its

still difficult, and so we use simulations.

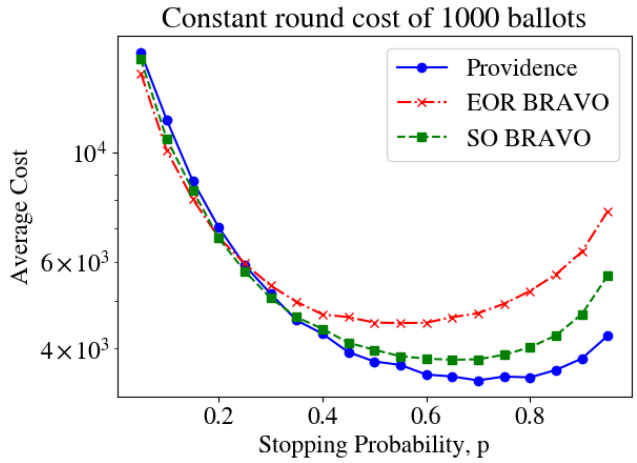

Figure 6: For cost parameters $c_b = 1$ and $c_r = 1000$, this plot shows the expected cost for various round schedule parameters p. Expected cost is found using Equation **??** and the average number of ballots and rounds in our simulations as the expected number of ballots and rounds.

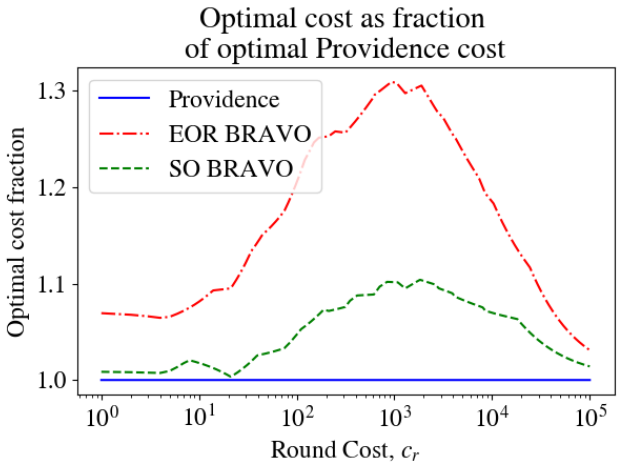

Figure 7: For varying round cost c_r, the optimal average cost achievable by each audit, as a fraction of the PROVIDENCE values. (We show the value for varying c_r since the value of c_r may differ significantly from case to case.) Note that values lower than 10^3 are probably unrealistic in US statewide contests; in Virginia, $c_r = 10^3$ corresponds to only about 15 minutes of persontime per county as the overhead per round.

corresponding stopping probability p. Figure 7 shows the optimal achievable workload for a wide range of per round costs. For very low round costs, the workload function approaches just the total number of ballots, and so, as seen in Figure **??**, the three approaches differ by less. On the other hand, for extremely larger values of round cost, the average number of ballots has little impact on the workload function, and so the three audits again have similar values. For more reasonable values of the round cost c_r, SO BRAVO and EoR BRAVO achieve minimum cost roughly 1.1 and 1.3 times greater than that of PROVIDENCE. Figure 8 shows the corresponding round schedule parameters p that achieve these minimal workloads. As expected, a overhead for each round means that larger round sizes are needed to achieve an optimal audit, and so for all three audit p increases as a function of c_r. Notice that PROVIDENCE is generally above and to the left of SO BRAVO, and SO BRAVO is generally above and to the left of EoR BRAVO. This relationship reflects the fact that for the same round cost, PROVIDENCE can get away with a larger stopping probability because it requires fewer ballots.

Precinct overhead. For a more complete model, we can also introduce container-level workload. The time to sample a ballot from an entirely new box is typically greater than to sample a ballot from an already-open box. Based on a Rhode Island pilot RLA report [4], this may mean that a ballot from a new container requires roughly twice the time as a ballot from an already-opened container. Typically available election results give per-precinct granularity of vote tallies, rather than individual container information. In Virginia, however, most precincts have a single ballot scanner whose one box

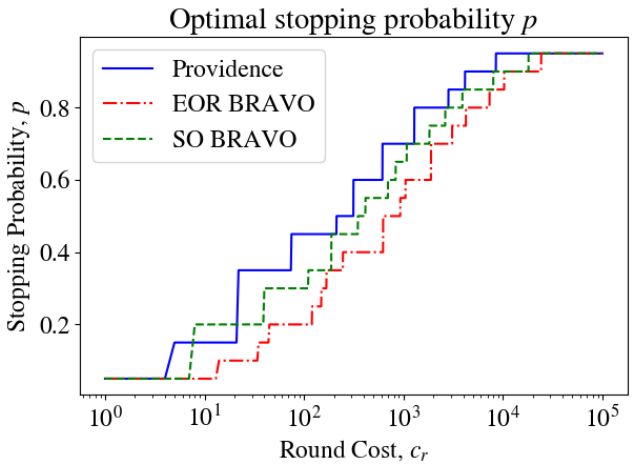

Figure 8: The optimal (cost-minimizing) stopping probability p for varying cost model parameters c_r. With $c_b = 1$, the varying value of c_r can equivalently be thought of as the ratio of the cost of a round to the cost of a ballot. (Note that the steps in this function are a consequence of our subsampling the workload function. That is, the workload-minimizing value of p for each c_r is only allowed to take on values at increments of 0.5.)

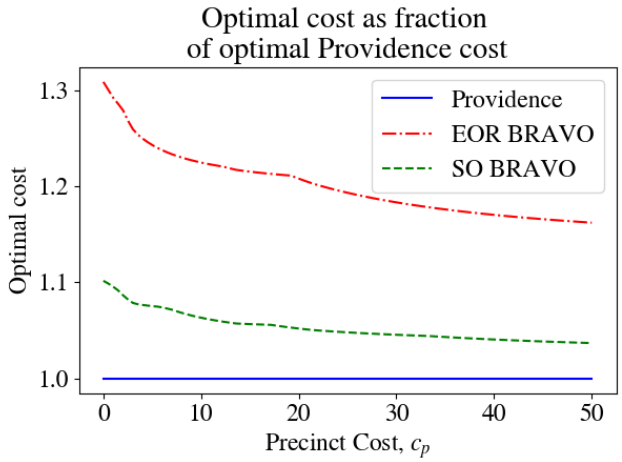

Figure 9: The optimal average cost of the audits we simulated using workload function 5 for varying c_p, given as a fraction of the value for PROVIDENCE

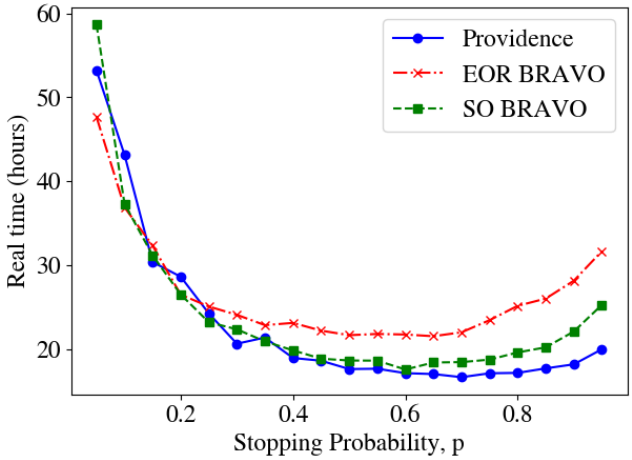

Figure 10: The real time as estimated by Equation 6 for varying p with expected values as estimated by our simulations.

has sufficient capacity for all the ballots cast in that precinct anyways, and so we model the per-container workload with an additional per-precinct constant cost, c_p. In this model, the workload estimate incurs an additional cost of c_p every time a precinct is sampled from for the first time in a round. That is, let E_{pi} be the expected number of distinct precincts sampled from in round i, and let $E_p = \sum_i E_{pi}$. Then the new model is

$$C(E_b, E_r, E_p) = E_b c_b + E_r c_r + E_p c_p \tag{5}$$

We can again explore the minimum achievable workloads under this model, as shown in Figure 9.

6.2 Real time

Given tight certification deadlines[2], the total real time to conduct the RLA is also an important factor to consider when planning audits. Because each county can sample ballots for the same round concurrently, the total real time for a round depends only on the slowest county. In Virginia, Fairfax County typically has the most votes cast by a significant difference; in the contest we consider, Fairfax County had 551 thousand votes cast, more than double the 203 thousand of second-highest Virginia Beach City. Consequently, we model the expected total real time T of an audit using just the largest county, and we define analagous variables for the expected values in just the largest county. For the largest county, let the expected total ballots sampled be $\bar{E}_b$, the expected number of rounds $\bar{E}_r$, and the expected number of distinct precinct samples summed over all rounds be $\bar{E}_c$. Similarly, we use real time per-ballot, per-round, and per-precinct cost variables, t_b, t_r, and t_p. So the real time of the audit is estimated by

$$T(\bar{E}_b, \bar{E}_r, \bar{E}_p) = \bar{E}_b t_b + \bar{E}_r t_r + \bar{E}_p t_p \tag{6}$$

As before, we can use our simulations to estimate $\bar{E}_b$, $\bar{E}_r$, and $\bar{E}_p$ using the corresponding averages over the trials. Available data to estimate values for t_b, t_r, and t_p is limited, and so we take as an example the values $t_b = 75$ seconds, $t_r = 3$ hours, and $t_p = 75$ seconds.[3] In practice, election officials could use our software and their own estimates of these values to explore choices for round schedules. Figure 10 shows how the esimated real time for these values differs as a function of p. It should be noted that real values of t_b, t_r, and t_p will vary greatly based on the number of parallel teams retreiving and checking ballots, the distribution of ballots and containers both in number and physical space, and other factors. We provide Figure 10 only as an example of the general shape and behavior of this function. Use of this optimal scheduling tool would depend on parameter estimates tailored to each case.

6.3 Misleading samples

Unfortunately, efficiency alone is not sufficient for planning audits. In the US today, election officials need to make legitimate safety considerations. When drawing a random sample of the ballots, it is always possible that the tally of the sample provides misleading information. In a random sample, a true loser may recieve more votes than the true winner. This

[2] Virginia recently passed legislation requiring pre-certification RLAs TODO check exactly.

[3] The value $t_b = 75$ seconds corresponds to a serial retrieval and interpretation of the ballots based on the [4] timing, $t_p = 75$ seconds corresponds to the approximate doubling in time for new-box ballots as reported in [4] in the ballot-level comparison timing data, and $t_r = 3$ hours is just a guess at an approximate order for this variable.

happens more often when the sample sizes are small. In these RLAs, a misleading sample in an early round is dealt with by drawing more ballots (moving on to another round), but in practice the implications of this approach may be dangerous.

Imagine that Alice beats Bob in an election contest both truly and by the announced results, but Bob's supporters are insistent he really won. When election officials carry out the RLA, they choose a small first round size in the hopes of achieving an efficient audit by getting to stop sooner. After the first round, by chance, there are more votes for Bob than for Alice in the sample. Bob's supporters celebrate their victory that the audit has in fact revealed that Bob really won, but the election officials have to explain that they are moving on to a second round. After the second round, there are more votes in the sample for Alice and sufficiently many that the risk limit is met and the audit now ends confirming the announced result that Alice won. This is an undesirable situation.

We introduce the notion of a *misleading sample*, any cumulative sample which, assuming the announced outcome is correct, contains more ballots for a loser than for the winner. We can again use our simulations to gain insight into the frequency of *misleading samples*. For each stopping probability p, Figure 11 gives the proportion of simulated audits that had a *misleading sample* at any point. Notably, this proportion is as high as 1 in 5 for the smaller stopping probability round schedules. Accordingly, we introduce a new parameter to our audit-planning tool, the maximum acceptable probability that the audit is misleading, the *misleading limit*.

In Figure 11, horizontal lines are included to show *misleading limits* of 0.1, 0.01, and 0.001. To achieve a probability of a misleading sample of at most 0.1, a round schedule with at least roughly $p = .3$ is needed. To achieve a probability of misleading of roughly 0.01, a round schedule with $p = 0.8$ is neeeded, and to achieve a probability of misleading of roughly 0.001, a round schedule with $p = 0.95$ is needed. It is not unreasonable to think that election officials might choose a *misleading limit* of 0.01, or smaller, given the state of public perception of election security in the US and the associated threats of violence. Consequently, the desired *misleading limit* may be a deciding constraint in the choice of round schedule.

If election officials wish to enforce a *misleading limit* for all the rounds, our simulation analysis could help. On the other hand, for a given round, it is straightforward to compute analytically the probability that a loser has more votes than the winner in the sample. Table 2 shows for various margins the minimum first round size n that guarantees a probability of a *misleading sample* at most $M \in \{0.1, 0.01, 0.001\}$. For all values of M and all margins, PROVIDENCE achieves a higher probability of stopping than either EoR BRAVO or SO BRAVO. As seen in the Table 2, for $M = 0.01$, the the minimum round sizes require at least roughly a 0.8 probability of stopping in the first round. Even if the most efficient audit schedule (by either workload or real time measures) would use a lower stopping probability p, the election officials may use this

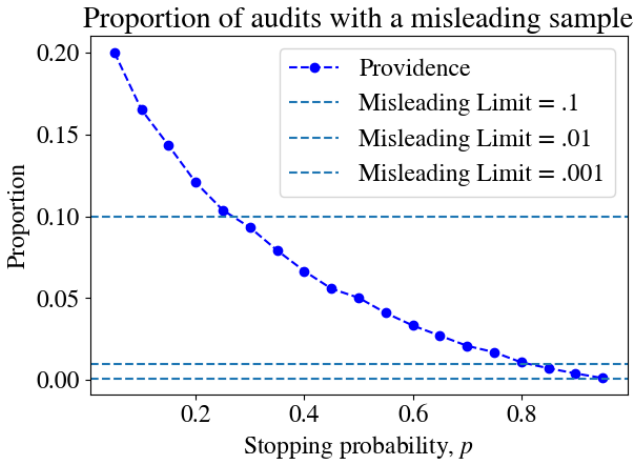

Figure 11: The proportion of simulated PROVIDENCE audits that had a *misleading sample* in any round.

constraint on the probability of a *misleading sample* as the deciding factor in selecting a first round size.

Other misleading RLA results.

7 Conclusion

A rigorous tabulation audit is an important part of a secure election. Ballot polling RLAs are commonly used and simple, not relying on special election equipment like comparison RLAs. We present PROVIDENCE which is the most efficient and secure ballot polling RLA, as efficient as MINERVA and flexible as BRAVO. We present proofs and simulation results to verify the claimed properties of PROVIDENCE, and we provide an open source implementation of the stopping condition and useful related functionality.

8 Availability

PROVIDENCE is implemented in the open source R2B2 software library for R2 and B2 audits. [10]

References

[1] Matthew Bernhard. *Election Security Is Harder Than You Think*. PhD thesis, University of Michigan, 2020.

[2] Michelle L. Blom, Peter J. Stuckey, and Vanessa J. Teague. Ballot-polling risk limiting audits for IRV elections. In Robert Krimmer, Melanie Volkamer, Véronique Cortier, Rajeev Goré, Manik Hapsara, Uwe Serdült, and David Duenas-Cid, editors, *Electronic Voting - Third International Joint Conference, E-Vote-ID 2018, Bregenz, Austria, October 2-5, 2018, Proceedings*, volume 11143

M	margin	n	Prov	SO	EoR
0.1	0.25	25	0.221	0.152	0.115
	0.2	41	0.178	0.169	0.105
	0.15	73	0.202	0.186	0.141
	0.1	163	0.222	0.182	0.107
	0.05	657	0.227	0.192	0.127
	0.04	1027	0.237	0.193	0.124
	0.03	1825	0.246	0.194	0.124
	0.02	4105	0.246	0.195	0.124
	0.01	16423	0.246	0.196	0.124
0.01	0.25	85	0.792	0.707	0.559
	0.2	133	0.826	0.71	0.593
	0.15	239	0.817	0.712	0.549
	0.1	539	0.805	0.717	0.567
	0.05	2163	0.817	0.721	0.569
	0.04	3381	0.82	0.722	0.563
	0.03	6011	0.824	0.723	0.573
	0.02	13527	0.824	0.723	0.57
	0.01	54117	0.824	0.724	0.57
0.001	0.25	149	0.962	0.889	0.783
	0.2	235	0.963	0.89	0.768
	0.15	421	0.958	0.894	0.801
	0.1	951	0.958	0.894	0.793
	0.05	3815	0.96	0.896	0.785
	0.04	5965	0.961	0.896	0.791
	0.03	10607	0.961	0.897	0.787
	0.02	23869	0.962	0.897	0.787
	0.01	95491	0.962	0.897	0.787

Table 2: For various margins, this table gives the minimum first round size n to achieve at most a probability M of a *misleading sample* in the first round. The corresponding stopping probabilities of PROVIDENCE, SO BRAVO, and EoR BRAVO are given for each value of n.

of *Lecture Notes in Computer Science*, pages 17–34. Springer, 2018.

[3] Oliver Broadrick, Sarah Morin, Grant McClearn, Neal McBurnett, Poorvi L. Vora, and Filip Zagórski. Simulations of ballot polling risk-limiting audits. In *Seventh Workshop on Advances in Secure Electronic Voting, in Association with Financial Crypto*, 2022.

[4] Common Cause, VerifiedVoting, and Brennan Center. Pilot implementation study of risk-limiting audit methods in the state of rhode island. https://www.brennancenter.org/sites/default/files/2019-09/Report-RI-Design-FINAL-WEB4.pdf.

[5] Zhuoqun Huang, Ronald L. Rivest, Philip B. Stark, Vanessa J. Teague, and Damjan Vukcevic. A unified evaluation of two-candidate ballot-polling election auditing methods. In Robert Krimmer, Melanie Volkamer, Bernhard Beckert, Ralf Küsters, Oksana Kulyk, David Duenas-Cid, and Mikhel Solvak, editors, *Electronic Voting - 5th International Joint Conference, E-Vote-ID 2020, Bregenz, Austria, October 6-9, 2020, Proceedings*, volume 12455 of *Lecture Notes in Computer Science*, pages 112–128. Springer, 2020.

[6] Mark Lindeman and Philip B Stark. A gentle introduction to risk-limiting audits. *IEEE Security & Privacy*, 10(5):42–49, 2012.

[7] Mark Lindeman, Philip B Stark, and Vincent S Yates. BRAVO: Ballot-polling risk-limiting audits to verify outcomes. In *EVT/WOTE*, 2012.

[8] Katherine McLaughlin and Philip B. Stark. Simulations of risk-limiting audit techniques and the effects of reducing batch size on the 2008 California House of Representatives elections. 2010.

[9] Katherine McLaughlin and Philip B. Stark. Workload estimates for risk-limiting audits of large contests. 2011.

[10] Sarah Morin and Grant McClearn. The R2B2 (Round-by-Round, Ballot-by-Ballot) library, https://github.com/gwexploratoryaudits/r2b2.

[11] Kellie Ottoboni, Matthew Bernhard, J. Alex Halderman, Ronald L Rivest, and Philip B. Stark. Bernoulli ballot polling: A manifest improvement for risk-limiting audits. *International Conference on Financial Cryptography and Data Security*, pages 226–241, 2019.

[12] Philip B. Stark. Simulating a ballot-polling audit with cards and dice. In *Multidisciplinary Conference on Election Auditing, MIT*, december 2018.

[13] Philip B. Stark and David A. Wagner. Evidence-based elections. *IEEE Secur. Priv.*, 10(5):33–41, 2012.

[14] VotingWorks. Arlo, https://voting.works/risk-limiting-audits/.

[15] Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

[16] Filip Zagórski, Grant McClearn, Sarah Morin, Neal McBurnett, and Poorvi L. Vora. The Athena class of risk-limiting ballot polling audits. *CoRR*, abs/2008.02315, 2020.

[17] Filip Zagórski, Grant McClearn, Sarah Morin, Neal McBurnett, and Poorvi L. Vora. Minerva– an efficient risk-limiting ballot polling audit. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 3059–3076. USENIX Association, August 2021.

A Proofs

Lemma 1. *For $0 < p_0 < p_a < 1$ and $n > 0$, the ratio $\sigma(k, p_a, p_0, n)$ is strictly increasing as a function of k for $0 \le k \le n$.*

Proof. From $0 < p_0 < p_a < 1$, we get

$$\frac{p_a}{p_0} > 1$$

and

$$1 - p_0 > 1 - p_a \implies \frac{1-p_0}{1-p_a} > 1,$$

and thus

$$\frac{p_a(1-p_0)}{p_0(1-p_a)} > 1.$$

Now simply observe that

$$\frac{p_a(1-p_0)}{p_0(1-p_a)} \cdot \sigma(k, p_a, p_0, n) = \frac{p_a(1-p_0)}{p_0(1-p_a)} \cdot \frac{p_a^k(1-p_a)^{n-k}}{p_a^k(1-p_a)^{n-k}}$$

$$= \frac{p_a^{k+1}(1-p_a)^{n-(k+1)}}{p_a^{k+1}(1-p_a)^{n-(k+1)}} = \sigma(k+1, p_a, p_0, n).$$

□

Lemma 2. *Given a monotone increasing sequence: $\frac{a_1}{b_1}, \frac{a_2}{b_2}, \ldots, \frac{a_n}{b_n}$, for $a_i, b_i > 0$, the sequence:*

$$z_i = \frac{\sum_{j=i}^n a_j}{\sum_{j=i}^n b_j}$$

is also monotone increasing.

Proof. Note that z_i is a weighted average of the values of $\frac{a_j}{b_j}$ for $j \ge i$:

$$z_i = \sum_{j=i}^n y_j \frac{a_j}{b_j}$$

for

$$y_j = \frac{b_j}{\sum_{j=i}^n b_j} > 0.$$

Further, $\sum_{j=i}^n y_j = 1$ and hence $y_j \le 1$ and $y_j = 1 \Leftrightarrow i = j = n$. Observe that, because $\frac{a_i}{b_i}$ is monotone increasing, $z_i \ge \frac{a_i}{b_i}$ with equality if and only if $i = n$. Suppose $i < n$. Then

$$z_{i+1} \ge \frac{a_{i+1}}{b_{i+1}} > \frac{a_i}{b_i},$$

and

$$z_i = y_i \frac{a_i}{b_i} + (1 - y_i) z_{i+1} < z_{i+1}.$$

Thus z_i is also monotone increasing. □

Lemma 3. *For $0 < p_0 < p_a < 1$ and $n > 0$, the ratio $\tau_1(k, p_a, p_0, n)$ is strictly increasing as a function k for $0 \le k \le n$.*

Proof. Apply Lemmas 1-2. □

Lemma 4. *Given a strictly monotone increasing sequence: $x_1, x_2, \ldots x_n$ and some constant A,*

$$A \le x_i \Leftrightarrow \exists i_{min} \le i \text{ s.t. } x_{i_{min}-1} < A \le x_{i_{min}} \le x_i,$$

unless $A \le x_1$, in which case $i_{min} = 1$.

Proof. Evident. □

Lemma 5. *For $\mathcal{A} = (\alpha, p_a, p_0, k_{j-1}, n_{j-1}, n_j)$-PROVIDENCE, there exists a $k_{min,j}$(PROVIDENCE, $p_a, p_0, k_{j-1}, n_{j-1}, n_j$) such that*

$$\mathcal{A}(X_j) = \textit{Correct} \iff k_j \ge k_{min,j}(\text{PROVIDENCE}, \boldsymbol{n_j}, p_a, p_0).$$

Proof. From Definition 4,

$$\mathcal{A}(X_j) = \text{Correct} \iff \omega_j(k_j, k_{j-1}, p_a, p_0, n_j, n_{j-1}) \ge \frac{1}{\alpha}.$$

Now to apply Lemma 4, it suffices to show that ω_j is monotone increasing with respect to k_j. For $j = 1$, we have $\omega_1 = \tau_1$, so ω_1 is strictly increasing by Lemma 3. For $j \ge 2$,

$$\omega_j(k_j, k_{j-1}, p_a, p_0, n_j, n_{j-1}, \alpha) =$$

$$\sigma(k_{j-1}, p_a, p_0, n_{j-1}) \cdot \tau_1(k_j - k_{j-1}, p_a, p_0, n_j - n_{j-1}).$$

As a function of k_j, σ is constant, and thus ω is strictly increasing by Lemma 3. Therefore by Lemma 4, we have the desired property. □

Lemma 6. *For $j \ge 1$,*

$$\frac{Pr[\boldsymbol{K_j} = \boldsymbol{k_j} \mid \boldsymbol{n_j}, H_a]}{Pr[\boldsymbol{K_j} = \boldsymbol{k_j} \mid \boldsymbol{n_j}, H_0]} = \sigma(k_j, p_a, p_0, n_j).$$

Proof. We induct on the number of rounds. For $j = 1$, we have

$$\frac{Pr[\boldsymbol{K_1} = \boldsymbol{k_1} \mid \boldsymbol{n_1}, H_a]}{Pr[\boldsymbol{K_1} = \boldsymbol{k_1} \mid \boldsymbol{n_1}, H_0]} = \frac{Pr[K_1 = k_1 \mid n_1, H_a]}{Pr[K_1 = k_1 \mid n_1, H_0]}$$

$$= \frac{\text{Bin}(k_1, n_1, p_a)}{\text{Bin}(k_1, n_1, p_0)} = \sigma(k_1, p_a, p_0, n_1).$$

Suppose the lemma is true for round $j = m$ with history $\boldsymbol{k_m}$. Observe that

$$\frac{Pr[\boldsymbol{K}_{m+1} = \boldsymbol{k}_{m+1} \mid \boldsymbol{n}_{m+1}, H_a]}{Pr[\boldsymbol{K}_{m+1} = \boldsymbol{k}_{m+1} \mid \boldsymbol{n}_{m+1}, H_0]}$$

$$= \frac{Pr[\boldsymbol{K}_m = \boldsymbol{k}_m \mid \boldsymbol{n}_{m+1}, H_a] \cdot Pr[K'_{m+1} = k'_{m+1} | \boldsymbol{k_m}, \boldsymbol{n}_{m+1}, H_a]}{Pr[\boldsymbol{K}_m = \boldsymbol{k}_m \mid \boldsymbol{n}_{m+1}, H_0] \cdot Pr[K'_{m+1} = k'_{m+1} | \boldsymbol{k_m}, \boldsymbol{n}_{m+1}, H_0]}$$

$$= \sigma(k_m, p_a, p_0, n_m) \cdot \frac{Pr[K'_{m+1} = k'_{m+1} | \boldsymbol{k_m}, \boldsymbol{n}_{m+1}, H_a]}{Pr[K'_{m+1} = k'_{m+1} | \boldsymbol{k_m}, \boldsymbol{n}_{m+1}, H_0]}$$

by the induction hypothesis. Then this is simply equal to

$$\sigma(k_m, p_a, p_0, n_m) \cdot \frac{\text{Bin}(k'_{m+1}, n'_{m+1}, p_a)}{\text{Bin}(k'_{m+1}, n'_{m+1}, p_0)}$$

$$= \frac{p_a^{k_m}(1-p_a)^{n_m - k_m}}{p_0^{k_m}(1-p_0)^{n_m - k_m}} \cdot \frac{p_a^{k'_{m+1}}(1-p_a)^{n'_{m+1} - k'_{m+1}}}{p_0^{k'_{m+1}}(1-p_0)^{n'_{m+1} - k'_{m+1}}}$$

$$= \sigma(k_{m+1}, p_a, p_0, n_{m+1})$$

□