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; this paper studies ballot polling RLAs which 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 Georgia and Pennsylvania, which are planning pre-certification ballot polling RLAs for the 2022 general election.

1 Introduction

It is well-known that electronic voting systems are vulnerable to software errors and manipulation which may be undetected. Errors and/or manipulation may not always change an election outcome, but we would want to know when they do. *Software independent* voting systems [12, 13] are ones where an undetected change in the software cannot lead to an undetected change in the election outcome. *Evidence-based elections* [15] use software independent systems to produce trustworthy evidence of outcome correctness; incorrect outcomes are detected with high probability when the evidence is examined. One approach to evidence-based elections is to use voter-verified paper ballots, store them securely, and perform public audits—a compliance audit to determine whether the ballots were stored securely; and a rigorous tabulation audit, known as a risk-limiting audit (RLA) [6], to determine whether the outcome is correctly computed from the stored ballots. Many US states have had pilot RLA programs. Additionally, some states allow RLAs to be used towards audit requirements, and some states require RLAs before elections can be certified.

We propose the PROVIDENCE audit, a new approach to the ballot polling RLA, and propose a new model for the work load of an election. We show that PROVIDENCE is superior to the popular ballot polling RLA BRAVO for real elections, and describe the use of our open source implementation by the Rhode Island Board of Elections for an audit of their 2021 elections. Our implementation of PROVIDENCE is likely to be useful later this year; ballot polling audits are expected to be used as pre-certification RLAs in at least one statewide contest in both Georgia and Pennsylvania for the 2022 general elections in the US.

1.1 Background

Ballot comparison RLAs require the fewest ballots of all known RLA approaches. On the other hand, they require the use of special election technology and are not always feasible. Ballot-polling RLAs require a larger number of ballots, but are more feasible because they do not require any additional functionality of the voting system. What is needed is a complete ballot manifest (a list of ballot storage containers and the number of ballots in each) which enables the creation of a well defined list of the ballots and their locations (the fifth ballot in box number 20, for example).

A ballot polling audit begins when a *round* [19] of multiple randomly-chosen ballots is drawn. A risk measure is then computed to determine whether (a) the audit ends in success (the election outcome is declared correct) or (b) another round should be drawn. Election officials would typically decide

to perform a full manual hand count if the audit does not stop in spite of drawing a large number of ballots, typically over multiple rounds. Ballot polling audits have been used in a number of US state pilots (California, Georgia, Indiana, Michigan, Ohio, Pennsylvania, Virginia and elsewhere) and in real statewide audits (Georgia, Virgina) [17] as well as in audits of smaller jurisdictions, such as Montgomery County, Ohio [19].

A *round-by-round (R2)* audit is one where the decision of whether to draw more ballots or not is taken after drawing a round of ballots; typically hundreds or thousands or tens of thousands of ballots in statewide elections. A *ballot-by-ballot (B2)* audit is the special case of round size one—when the decision is made after each ballot is drawn. The popular BRAVO audit requires the smallest expected number of ballots when the announced tally of the election is correct, and stopping decisions are taken a ballot at a time (that is, when it is used as a B2 audit). Election officials typically draw ballots in large round sizes, and BRAVO needs to be modified for use in this manner. For use as an R2 audit, the BRAVO stopping condition can be applied once at the end of each round (End-of Round (EoR)), or retroactively after each ballot drawn if ballot order is retained (Selection-Ordered (SO)). SO BRAVO is closer to the original B2 BRAVO, and requires fewer ballots on average than EoR BRAVO. But it requires the additional effort of tracking the order of ballots.

Zagórski *et al.* propose ballot polling RLA MINERVA [19], which does not need ballot order and relies only on sample and round tallies. They prove that it requires fewer ballots than EoR BRAVO when both audits have the same pre-determined round schedule and the true tally is as announced. They also present first-round simulations demonstrating that MINERVA draws fewer ballots than SO BRAVO in the first round for large first round sizes when the true tally is as announced. Broadrick *et al.* provide further simulations that show MINERVA requires fewer ballots over multiple rounds and for lower stopping probability [3], though the improvement from using MINERVA over either version of BRAVO decreases with round size.

A major limitation of MINERVA is that one needs to determine the round schedule before the audit begins, because MINERVA has not been shown to be risk-limiting if an adversary can choose subsequent round sizes after knowing the sample drawn. BRAVO, on the other hand, is not limited in this manner. This allows BRAVO audits the flexibility of choosing smaller subsequent round sizes if the sample drawn so far is a "good" sample. An open question is whether a ballot polling RLA exists with the efficiency of MINERVA and this flexibility of BRAVO.

A major limitation of our understanding of the ballot polling problem as a community is that we use the number of ballots drawn or values proportional to this number [1, 4, 9] as measures of the workload of an audit. If this were a correct measure of the workload of an audit, we would want to use B2 audits (round size is one) and make decisions about stopping the audit after drawing each ballot, because this leads to the smallest expected number of ballots. Election officials, on the other hand, greatly prefer drawing many ballots at once. This preference is likely because each round has an overhead workload as well, including setting up the round and communicating among the various localities involved in conducting the audit (for example, audits of statewide contests involve the drawing of ballots at county offices where the ballots are stored). Further, there is an overhead to finding a storage box and unsealing it. For large round sizes, multiple ballots may be drawn at once from a box, and the number of boxes retrieved is smaller than the number of ballots (storage boxes commonly contain many hundreds of ballots each). For smaller round sizes, the number of times a box is retrieved would be roughly identical to the number of ballots drawn, as it is unlikely that a single box will hold multiple ballots from the sample. Finally, in the current environment of misinformation, election officials would want to ensure that the probability of a misleading audit sample (falsely indicating that the loser won) is very small, which implies that round sizes should be large. Thus the workload of an audit is not simply a linear (or affine) function of the number of ballots drawn. Relatedly, an optimal round schedule is not completely determined by the expected number of ballots drawn. It depends on other variables as well. The consideration of all these variables is necessary while planning an audit.

1.2 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 uses number of ballots similar to those of MINERVA, both fewer than either version of BRAVO.
3. Results and analysis from the use of PROVIDENCE in a pilot audit in Rhode Island.
4. Open source implementation of PROVIDENCE.
5. A model of workload that includes the overhead effort of each round and the overhead effort of retrieving a storage unit of ballots; simulations that illustrate the use of this model to compare the different types of ballot polling audits and to plan an audit with minimal workload.
6. An analysis of round size as a function of the maximum acceptable probability of a misleading audit sample.

Our results demonstrate the superiority of PROVIDENCE over the other audits. Our work may be used by election officials to plan ballot polling audits, including in Georgia and Pennsylvania in 2022.

1.3 Organization

Section 2 describes related work. Section 3 describes the PROVIDENCE audit, section 4 the simulations comparing the number of ballots drawn using various ballot polling audits and section 5 the use of PROVIDENCE in an audit carried out by the Board of Elections of Rhode Island. Section 6 presents our workload model and describes its use for a ballot polling audit using details of the 2020 US Presidential election in the state of Virginia. Our conclusions, the availability of an audit implementation and acknowledgements may be found in sections 7, 8 and 9 respectively.

2 Related work

The BRAVO audit [7] is the most popular ballot polling audit. When ballots are sampled one at a time, it is the audit with the smallest expected number of ballots drawn.

The MINERVA audit [18, 19] was developed for use with large first round sizes, and has been proven to be risk limiting when the round schedule for the audit is fixed before any ballots are drawn. First-round sizes for a stopping probability of 0.9 when the announced tally is correct have been shown to be smaller than those for EoR and SO BRAVO for a wide range of margins; simulations [18] support these observations. 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 provide a means of educating the public and election officials [14] and to understand audit properties [2, 5, 8, 9]. There is work measuring the amount of time taken to examine a single ballot [4]. Simple workload estimates may be obtained by using the number of ballots drawn [11], a more thorough workload estimation model includes the time taken to access individual ballots [1].

We now summarize the model drawing largely from the notation and terminology of [3, 7, 18, 19]. The model is related work and not claimed to be original to this work.

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 tied election and the alternative hypothesis H_a to an election tally as announced. (When the number of ballots is odd, H_0 corresponds to the announced loser winning by one ballot.) Thus the null hypothesis is the outcome distinct from the announced one which is most difficult to detect; the probability of failing to detect it, given that the null hypothesis is true, is the worst case such probability and should be below the risk limit [16].

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

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

It is easy to see that the ratio σ is 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 cumulative round sizes and the corresponding sequence of cumulative 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 the ratio of the values of the probability distribution functions, MINERVA uses the ratio of their *tails*.

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 each round. The proportion of ballots for the reported winner under the alternative hypothesis and null hypothesis are p_a and p_0 respectively.*

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

3 PROVIDENCE

In this section we motivate and introduce the stopping condition of PROVIDENCE.

The MINERVA stopping condition tests the tail ratios of the probability distributions (given the null and alternate hypotheses) of winner votes K_j in round j. Note that these distributions are not conditioned on k_{j-1}. Thus they may be viewed

as a weighted average of various distributions, each conditioned on a possible value of K_{j-1}. The proof of MINERVA's risk-limiting property relies on the fact that, at every stopping point, the tail of the probability distribution function given H_0 is at most α times that given H_a. However, because these distributions are averaged over the various values of K_{j-1}, this approach to the proof works only if the next round size is the same for all values of K_{j-1}. If not, this average pdf will not correctly represent the pdf of K_j. For example, if some values of K_{j-1} would result in larger next round sizes, the pdf for these values of K_{j-1} should not contribute to the average distribution function of K_j for smaller round sizes.

For the PROVIDENCE audit we test the tail ratios of the distributions conditioned on k_{j-1} and multiplied by the weighting factor. That is, we test this tail ratio separately for each weighted distribution corresponding to each possible value of K_{j-1}. Because of this, we no longer require that all values of k_{j-1} should result in the same next round size. This results in a great deal of flexibility in multiple round audits.

3.1 Definition

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:*

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

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

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

Notice that for $j \geq 2$, unlike τ_j, computing ω_j requires no convolution and is hence considerably more computationally efficient.

3.2 Proof of Risk-Limiting Property

We now 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

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

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

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

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

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

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

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

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

□

3.3 Resistance to an adversary choosing round sizes

Filip TBD

4 Simulations

We use simulations to provide additional evidence for theoretical claims and gain insight into audit behavior. As in [3], 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, by design, have a low probability of stopping and

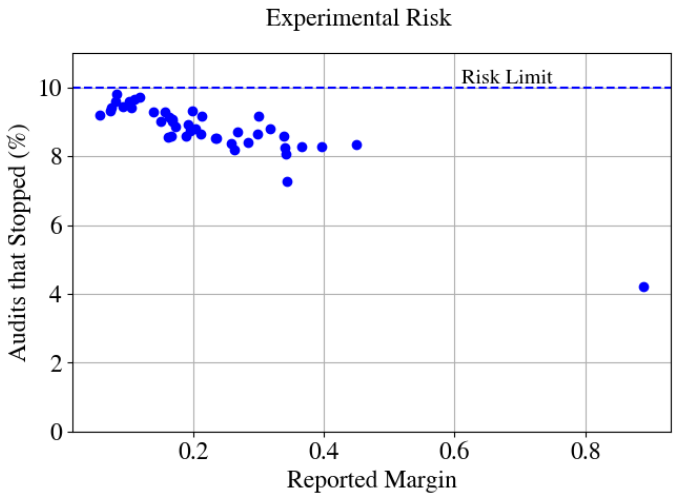

Figure 1: The fraction of simulated PROVIDENCE audits on tied elections that stopped in any rounds (we performed five rounds at a 10% risk limit). This value is an estimate of the maximum risk of the PROVIDENCE audit.

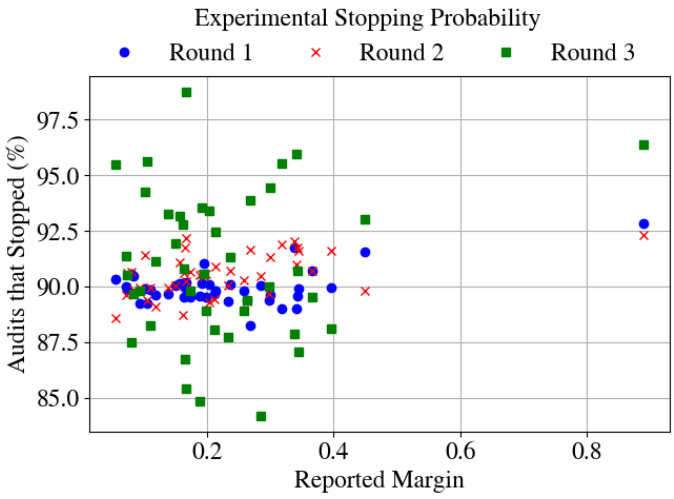

Figure 2: The fraction of simulated PROVIDENCE audits of the election as announced that stopped for each round. This value is an estimate of the stopping probability conditioned on the sample of the previous round. The average fraction for rounds 1, 2, and 3 is 89.96%, 90.52%, and 90.98% respectively. We show only the first three rounds since so few audits make it to rounds 4 and 5.

hence require all rounds and quickly increase in sample size. We use the simulator in the R2B2 software library [10]. For each margin, for each hypothesis, we perform 10^4 trials. All trials use a 10% risk limit, a maximum of 5 rounds, and a conditional stopping probability of 0.90 in each round. That is, each subsequent round size is selected to be large enough, assuming the announced tally is correct and given the tally of previous rounds, to give a 0.90 probability of stopping in the current round.

In the simulations of PROVIDENCE audits of a tied election, the fraction of audits that stop, as shown in Figure 1, is an estimate of maximum risk. For all margins, this estimated maximum risk is less than the risk limit, supporting the claim that PROVIDENCE is risk-limiting.

Simulations of audits of the election as announced provide insight into stopping probability and number of ballots drawn when the election is as announced. We wish the stopping probability to be as predicted, and the number of ballots drawn to be small. Figure 2 shows that the stopping 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 plots the probability of stopping as a function of the 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. 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 correspond to many days of work.

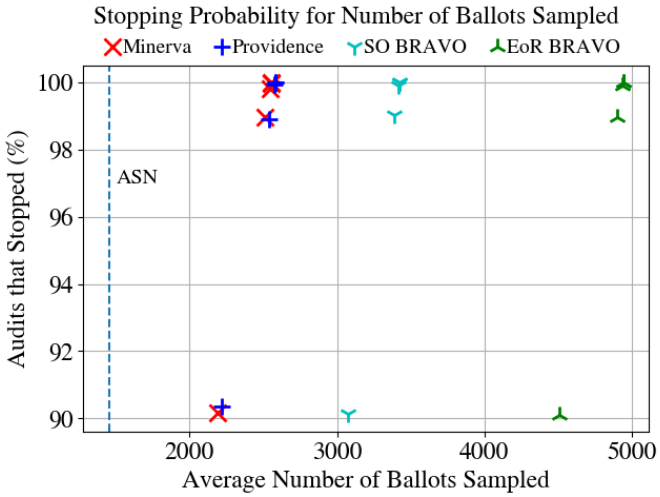

Figure 3: For the entire audit, consisting of all five rounds, the estimated stopping probability for average number of ballots drawn for PROVIDENCE, MINERVA, EoR BRAVO, and SO BRAVO.

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.

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 in the November 202 1 election: Portsmouth's Issue 1, "School Construction and Renovation Projects". The question had announced margin 25.67% and 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.

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: the true winner has fewer votes in the audit sample than some other candidate. 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 maximum allowed probability of a misleading audit sample.

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) WHAT ARE Eb AND Er? 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 have 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 trial audits for 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.

[1] For this particular round schedule scheme, computing the expected number of rounds is possible analytically, but the expected number of ballots is still difficult, and so we use simulations.

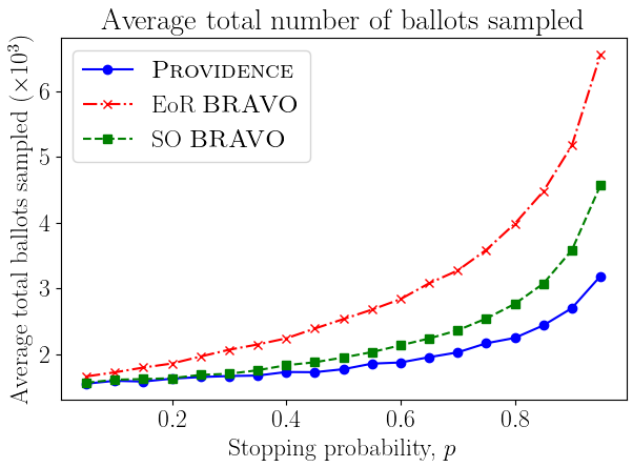

Figure 4: The total number of ballots sampled on average in our simulations for various round schedules parameterized by p the conditional stopping probability used to select each round size.

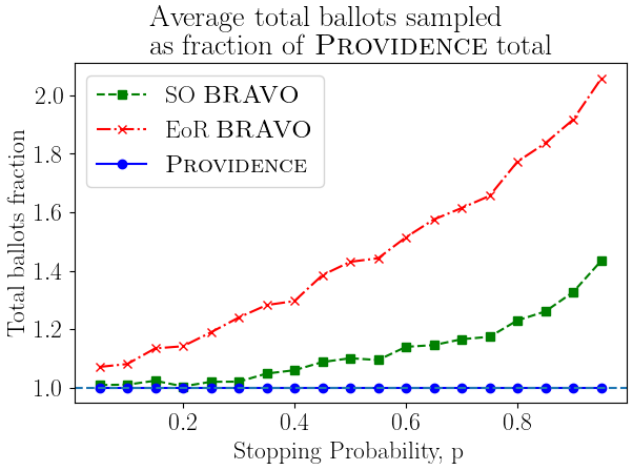

Figure 5: The total number of ballots sampled on average in our simulations given as a fraction of those sampled by Providence, for various round schedules parameterized by p the conditional stopping probability used to select each round size.

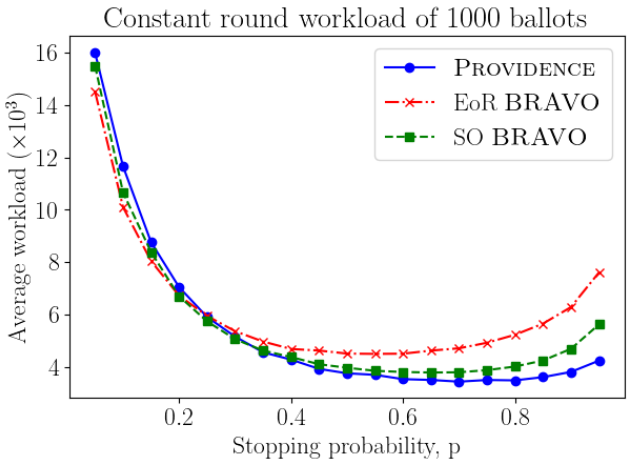

Figure 6: For workload parameters $w_b = 1$ and $w_r = 1000$, this plot shows the expected workload for various round schedule parameters p. Expected workload is found using Equation **??** and the average number of ballots and rounds in our simulations as the expected number of ballots and rounds.

Consequently, we now consider a model with a constant per-ballot workload w_b and a constant per-round workload w_r. So for an audit with expected number of ballots E_b and expected number of rounds E_r, we estimate that the workload W of the audit is

$$W(E_b, E_r) = E_b w_b + E_r w_r + C \tag{4}$$

Note there is also some constant overhead of workload for the whole audit, namely C in Equation 4, which we take to be zero in our examples but could be used by election officials for their estimates. For simplicity, (and without loss of generality), we assume the per ballot workload is one, $w_b = 1$. Then the per round workload w_r tells us the workload of a round as a number of ballots. We begin with $w_r = 1000$ 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 $w_r = 1000$ 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 workloads are achieved by selecting higher stopping probability; Providence achieves the lowest minimum average workload is achieved at roughly 0.7.

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

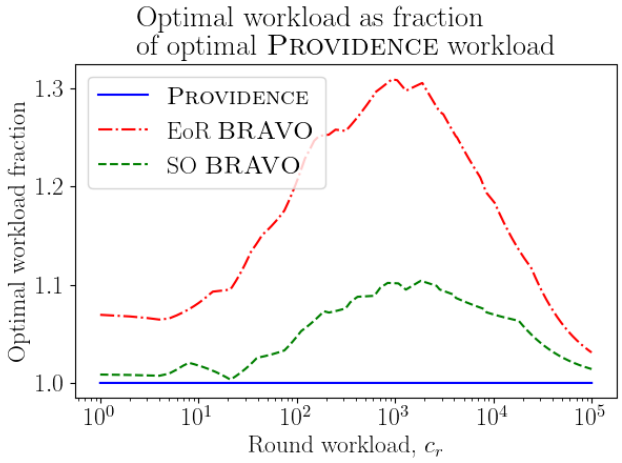

Figure 7: For varying round workload w_r, the optimal average workload achievable by each audit, as a fraction of the PROVIDENCE values. (We show the value for varying w_r since the value of w_r may differ significantly from case to case.) Note that values lower than 10^3 are probably unrealistic in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

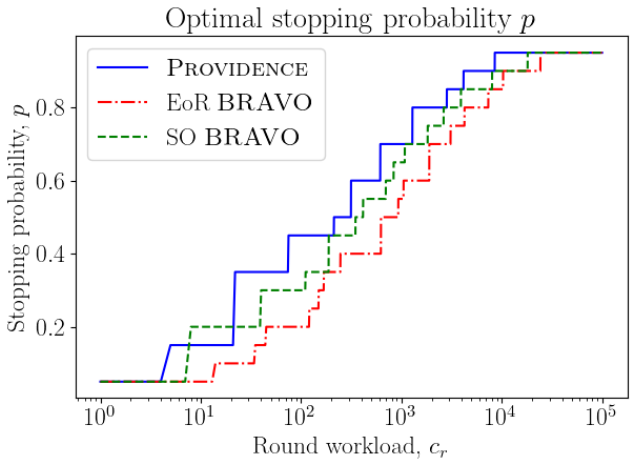

Figure 8: The optimal (workload-minimizing) stopping probability p for varying workload model parameters w_r. With $w_b = 1$, the varying value of w_r can equivalently be thought of as the ratio of the workload of a round to the workload of a ballot. (Note that the steps in this function are a consequence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

audits and its corresponding stopping probability p. Figure 7 shows the optimal achievable workload for a wide range of per round workloads. For very low round workloads, 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 workload, 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 workload w_r, SO BRAVO and EoR BRAVO achieve minimum workload 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 w_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 workload, 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 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 workload, w_p. In this model, the workload estimate incurs an additional workload of w_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

$$W(E_b, E_r, E_p) = E_b w_b + E_r w_r + E_p w_p + C \qquad (5)$$

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;

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

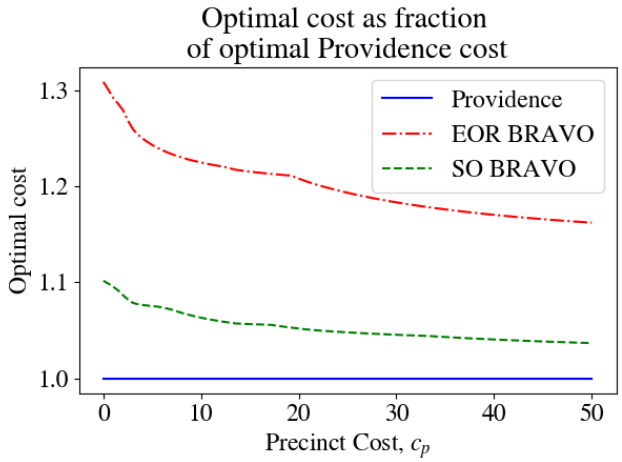

Figure 9: The optimal average workload of the audits we simulated using workload function 5 for varying w_p, given as a fraction of the value for PROVIDENCE

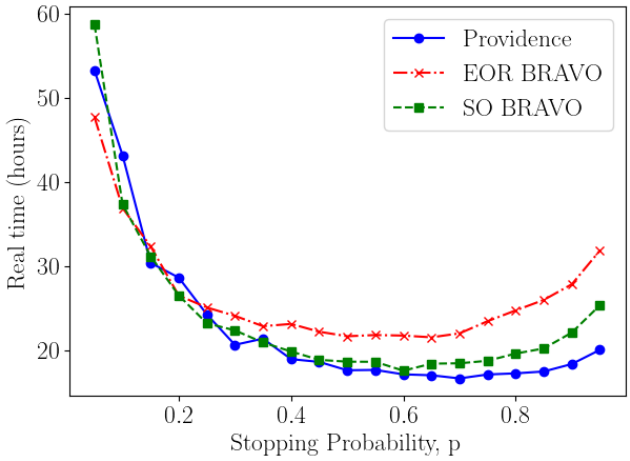

Figure 10: The real time as estimated by Equation 6 for varying p with expected values as estimated by our simulations.

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 workload 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.

[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.

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 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

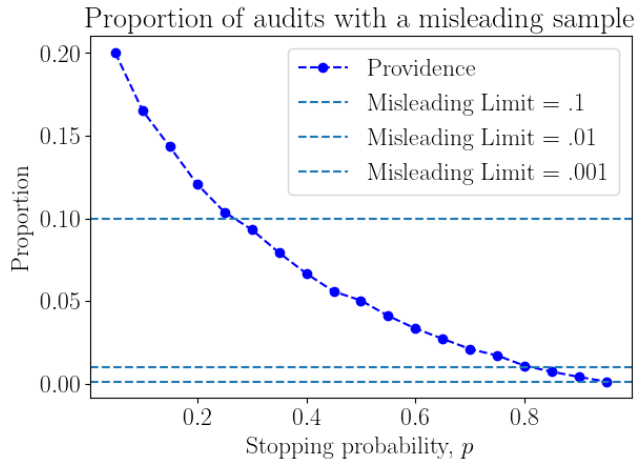

Figure 11: The proportion of simulated PROVIDENCE audits that had a *misleading sample* in any round.

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, to enforce $M = 0.01$ requires minimum round sizes with 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 to choose the first round size, the election officials may opt to use this constraint on the probability of a *misleading sample* as the deciding factor in planning their audits.

Misleading SO Risk Measures. As we consider the idea of misleading samples, it is noteworthy that SO BRAVO suffers

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.

from a different and unique type of misleading result.

After drawing $n > 1$ ballots in a round, some number k of them are votes for the announced winner. There are $\binom{n}{k}$ possible sequences of ballots which can lead to such a sample. Given the value of k, however, the particular sequence of the sample contains no additional information about whether the sample is more likely under one the alternative or null hypotheses. That said, an SO BRAVO RLA will not stop as a function of n and k but also as a function of the sequence. In particular, if the sequence of ballots is such that the standard BRAVO stopping condition was met for some $n' < n$ and corresponding $k' < n$, the audit will stop, even if by the end of the sequence the values k and n no longer meet the BRAVO condition. To be clear, this is not a mathematical issue; stopping in such cases is still a correct application of Wald's SPRT result. The misleading nature of such stoppages is the note we are making.

It is easy to use our simulations to see how often this occurs.

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]

9 Acknowledgements

The authors are grateful to the Rhode Island Board of Elections for conducting a pilot PROVIDENCE RLA.

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 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. NSF report, 2010.

[9] Katherine McLaughlin and Philip B. Stark. Workload estimates for risk-limiting audits of large contests. Honors Thesis, University of California, Berkeley, 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] Ronald L Rivest. On the notion of "software independence" in voting systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1881):3759–3767, 2008.

[13] Ronald L. Rivest and John P. Wack. On the notion of "software independence" in voting systems. Prepared for the TGDC, and posted by NIST at the given url.

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

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

[16] Poorvi L. Vora. Risk-limiting Bayesian polling audits for two candidate elections. *CoRR*, abs/1902.00999, 2019.

[17] Verified Voting. Audit law database, https://verifiedvoting.org/auditlaws/.

[18] 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.

[19] 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 $j \geq 1$,*

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

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

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

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

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

□