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

immediate

Abstract

A Risk-Limiting Audit (RLA) is a statistical election tabulation audit with a rigorous error guarantee. We present ballot polling RLA 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. We quantify the problem of obtaining a misleading audit sample when rounds are too small, demonstrating the importance of the resulting constraint on audit planning. We present simulation results demonstrating the superiority of PROVIDENCE using these measures and describing an approach to planning audit round schedules.

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 has been integrated as an option in Arlo, the most commonly used RLA software.

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 [15, 16] are ones where an undetected change in the software cannot lead to an undetected change in the election outcome. *Evidence-based elections* [18] 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) [10], to determine whether the outcome is correctly computed from the stored ballots. A *risk-limiting audit* guarantees a minimum probability of a full hand count if the election outcome is incorrect. Conversely, it guarantees a maximum probability of audit error, termed the *risk limit*, which is the maximum probability with which the audit would declare an incorrect election outcome as being correct.

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.

1.1 Background on RLAs

We provide here the background necessary to evaluate our contributions.

All RLAs sample one or more ballots at random; we will refer to each such set of ballots as a *round*. The ballots are manually examined, and a stopping condition computed, which determines whether (a) the audit ends in success (the election outcome is declared correct) or (b) another round should be drawn (more information is needed before a determination). In principle, the stopping condition could indicate a third option too: (c) the audit proceeds to a full manual hand count. However, a manual hand count presents significant logistical challenges, and there is always a chance that it will have been unnecessary, hence (c) is generally not incorporated. 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. They would be influenced by the certification deadline, the estimated number of human hours required for another round, the logistical costs of a full hand count, and the impact of any decision on citizen confidence.

1.1.1 Types of RLAs

In a ballot comparison RLA [10], the manual interpretation of each sampled ballot is compared to the corresponding Cast Vote Record (CVR), which is the machine interpretation of

the ballot. Ballot comparison RLAs require the fewest ballots of all known RLA approaches, but also require a means of identifying the CVR corresponding to a particular ballot. A typical approach is to use a ballot serial number on both paper ballot and CVR. When voters vote in precincts, however, serial numbers on ballots can enable the correlation of ballots with voters, and ballots are typically not numbered. Additionally, some voting systems do not record a CVR. One may perform a transitive audit by rescanning unnumbered voted ballots with special scanners which produce CVRs and also print numbers on the ballots as they are scanned. This requires an investment in a sufficiently large number of such printers and in the human effort of rescanning ballots, and is not always feasible.

In a ballot polling RLA [10], the manual interpretations of the sampled ballots are simply tallied. Ballot polling RLAs require a much larger number of ballots than ballot comparison RLAs, but are more feasible because they do not require any additional functionality of the voting system, and, in particular, do not need CVRs. 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 batch comparison RLA [7] samples batches of ballots (typically, a batch is a storage box of ballots) and compares the manual tally of each sampled batch with the announced tally of that batch. Thus, while this type of audit does not need CVRs, it does need both a ballot manifest and a public declaration of the tally of each batch. This approach typically requires the sampling of a very large number of ballots. However, the process is most similar to that which election officials already use when they perform fixed-percentage post-election audits, where, for example, 2% of the batches are manually tallied. (Note that, for an RLA, the number of batches tallied is not fixed; the risk limit is. A smaller number of batches might be sufficient, or a larger number necessary, for an audit with the required risk limit.)

This paper focuses on ballot polling RLAs which 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) [20] as well as in audits of smaller jurisdictions, such as Montgomery County, Ohio [24].

1.1.2 Ballot Polling Audits

Ballot polling audits proceed as follows.

1. A first round [24] size—the number of ballots sampled before first checking the stopping condition—is chosen.

2. Ballots on the ballot manifest are sampled uniformly at random using a pseudorandom number generator typically seeded by a natural source of randomness like rolling dice.

3. The physical ballots are found and manually interpreted, recording the manual interpretations.

4. Based on the manual interpretations, the stopping condition is computed.

5. If more ballots are to be drawn, the next round size is chosen. Round sizes, including the first one, may be computed based on a desired probability of audit completion at the end of the round, and may take into consideration loose estimates of the resources required. For RLAs required by statute or law, certification deadlines would play a large role, as the audit would need to be completed before the deadline.

A *round-by-round (R2)* audit is the general audit, 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: see for example [**?**, 7], and note that, in addition to allowing users to directly enter a round size, Arlo provides choices of stopping probabilities of 0.9, 0.8 and 0.7, and the expected number of ballots required by BRAVO. Further, at both audits we attended, election officials chose stopping probabilities of 0.9 and we are not aware of any ballot polling RLA performed on ballots cast in a governmental election that drew ballots one at a time (though the stopping condition can be computed one ballot at a time, the ballots are drawn in rounds). BRAVO hence cannot be used as a B2 audit in these scenarios.

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.

1.1.3 Adversarial Model for RLAs

Detailed descriptions of best practices for post-election audits may be found in [5, 8]. For our purposes, we will assume that the best practices are followed: the paper trail consists of hand-marked paper ballots and is secured; a public compliance audit is carried out before the RLA to ensure that the processes for securing the paper trail were followed; voter authentication and registration processes were verified, and only legitimate voters cast no more than a single vote each; the risk-limiting

audit is public. We will further assume that all software used in the RLA is open source and well-defined, so its output may be reproduced and thus verified by an observer wishing to do so with their own software.

Referring to the ballot polling audit steps described above, we further assume that a secure PRNG is used; the seed is generated uniformly at random in a public process; the process of locating ballots is publicly observable and the located ballots can be viewed by the public. Because the PRNG is well-defined, as is the stopping condition, we may assume that the stopping condition is correctly computed from knowledge of the seed and the drawn ballots. Thus the only variable is round size. We define a *weak adversary* as one who can choose the first round size and a *strong adversary* as one who can choose any round size.

1.2 The Literature and Open Questions

Zagórski *et al.* propose ballot polling RLA MINERVA [24], which does not need ballot order and relies only on sample and round tallies. They prove that it is risk-limiting when the number of relevant ballots drawn in each round is pre-determined before any ballots are examined; that is, for a weak adversary. They do not address the case of a strong adversary (such as an audit insider) who can determine the size of the next round after knowing what votes are on the ballots sampled thus far. In particular, an open question about MINERVA is whether the computation of a risk limit assuming a weak adversary applies to an attack by a strong adversary, or is the risk limit computation incorrect when the adversary is strong? Can the strong adversary increase the audit's error probability beyond its declared risk limit? Or is there no probabilistic adversarial advantage to being able to compute next round sizes after knowing the drawn sample? We do not answer this question, and to our knowledge, it remains open.

Until MINERVA is proven to be risk-limiting to a given risk limit for the strong adversary, it may not be used in audits whose round sizes are not pre-determined. This presents a major limitation, because the stopping probability of the next round is better estimated using information of the sample drawn thus far, but this would not be allowed for MINERVA. The current implementation of MINERVA integrated as an option in Arlo uses a fixed multiplier of the current round size to compute the next round size, thus allowing the first round to be computed as desired, and fixing the next round sizes thereafter. Note that every draw may contain invalid or irrelevant ballots, and thus the true number of relevant ballots can never be predetermined. However, because this is random, and not controllable by an adversary once the size of the draw is fixed, we assume that a fixed draw size is sufficient to limit adversaries to weak ones, though this is not explicitly proven in [24].

Zagórski *et al.* 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 [6], though the improvement from using MINERVA over either version of BRAVO decreases with round size.

The risk limit for B2, EoR and SO BRAVO is fixed whether the adversary is strong or weak. 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 [2, 7, 13] 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. As described above, election officials, on the other hand, greatly prefer drawing many ballots at once. This preference is likely due to the following.

Firstly, 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).

Secondly, 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.

1.3 Our Contributions

Our primary contribution is a new RLA, PROVIDENCE, which gives the efficiency of MINERVA and is also resistant to a strong adversary. The stopping condition for MINERVA does not take into account the sample obtained in previous rounds, and, in [24], its risk limit is estimated through weighted averages across multiple rounds, assuming that round sizes do not depend on the previous sample. We are able to derive a new stopping condition for which a far simpler proof of the risk-limiting property is possible. In particular, this proof does not require an assumption about round sizes. We provide the following:

1. Proof that PROVIDENCE is an RLA and resistant to a strong adversary.
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. 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.
5. An analysis of round size as a function of the maximum acceptable probability of a misleading audit sample.
6. Open source implementation of PROVIDENCE and audit planning tools.

1.4 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

Bernhard provides a good description of the RLA and its assumptions, and also describes the process on the ground [3].

The BRAVO audit [11] 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 [23, 24] 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 [23] support these observations. Additional simulations [6] 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 [17] and to understand audit properties [4, 9, 12, 13]. There is work measuring the amount of time taken to examine a single ballot [7]. Simple workload estimates may be obtained by using the number of ballots drawn [14], a more thorough workload estimation model includes the time taken to access individual ballots [2].

We now summarize the model drawing largely from the notation and terminology of [6, 11, 23, 24]. 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 [19].

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 K is the random variable indicating the number of ballots in the sample that contain a vote for the reported winner, it is easy to see that the ratio σ is the likelihood ratio:

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

BRAVO is an instance of Wald's Sequential Probability Ratio Test (SPRT) [22]. In the more general SPRT, the test can also reject the alternative hypothesis, and there is an additional parameter β, the probability of incorrectly rejecting the alternative hypothesis. In an RLA, this corresponds to the audit having a third output: *proceed to a full manual count of the ballots*. In the existing literature, ballot polling audits do not include this possibility (i.e. they set $\beta = 0$) in order to give maximum flexibility to election officials in choosing when to proceed to a full manual count.

Where BRAVO uses the ratio of the values of the probability distribution functions, MINERVA uses the ratio of their *tails*. Now it becomes useful to have shorthand for a sequence of cumulative round sizes and the corresponding sequence of cumulative winner ballot tallies. We use:

$$\boldsymbol{n_j} \triangleq (n_1,n_2,\ldots,n_j) \quad \text{and} \quad \boldsymbol{k_j} \triangleq (k_1,k_2,\ldots,k_j)$$

Also, let K_j be the random variable indicating the cumulative number of ballots in the sample after the jth round is drawn.

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.*

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

In this work, we consider ballot polling RLAs only and thus compare PROVIDENCE with BRAVO and MINERVA.

3 PROVIDENCE

In this section we introduce the stopping condition of PROVIDENCE and prove some properties.

Recall that the proof that the MINERVA audit is risk-limiting assumes that the round schedule of MINERVA is predetermined and that, in particular, an adversarial auditor cannot determine the next round size after drawing a sample. This presents difficulties because a non-adversarial election official might want to draw a small next round if the current sample comes close to satisfying the risk limit. Because the MINERVA round size is predetermined, however, the election official would be required to draw a larger round size than necessary for the sample. Conversely, if the current sample is not at all close to satisfying the risk limit, it would be advantageous to draw a larger round than the predetermined round size.

We now formalize these notions of weak and strong adversaries.

Definition 4 (Weak Adversary)**.** *A weak adversary may choose the first round size as a function of audit parameters. That is, the first round size is a function*

$$n_1(\alpha,p_a,p_0).$$

Definition 5 (Strong Adversary)**.** *A strong adversary may choose any round size as a function of audit parameters and all precreding samples. That is, the first round size is a function*

$$n_1(\alpha,p_a,p_0),$$

and for all rounds $j \geq 2$, the round size is a function

$$n_{j+1}(\alpha,p_a,p_0,\boldsymbol{k}_{j-1},\boldsymbol{n}_{j-1}).$$

Before defining PROVIDENCE, we give some intuition for how it is designed to avoid the problem of MINERVA. In round j, for every possible value of K_j for which the audit stops, a risk is incurred. To obtain the total risk for the round, one adds the risks corresponding to each value of K_j for which the audit can stop, weighted by the probability of drawing that value of K_j. The stopping condition provides relationships among various quantities.

In MINERVA, the stopping condition relates the weighted average of the risks to the weighted average of the stopping probabilities over all values of K_j, for a given round size. Separate relationships between risk and stopping probability are not available for individual values of K_j. If the next round size depends on K_j, we do not have expressions relating the risks, and so MINERVA may be vulnerable to an adversary choosing round sizes [23, 24].

In PROVIDENCE, we choose a stopping condition that applies separately to the risk and stopping probabilities for each value of K_j, avoiding the problem of MINERVA, and allowing for optimal round size choices, which depend on the drawn sample. The PROVIDENCE audit is risk-limiting even if an adversarial auditor determines round sizes after drawing the sample, and next round size computations may use knowledge of the current sample.

3.1 Definition

Definition 6 ($(\alpha,p_a,p_0,k_{j-1},n_{j-1},n_j)$-PROVIDENCE)**.** *For cumulative round size n_j for round j and a cumulative k_j ballots for the reported winner found in round j, 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:*

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

Notice that PROVIDENCE requires the computation of τ_j for $j = 1$ and no other values of j. The value of τ_1 is simply the ratio of the the tails of the binomial distributions for the two hypotheses and can be fairly efficiently computed. The computation of τ_j for $j \geq 2$, as required in MINERVA, relies on the convolution of two probability distribution functions and is hence computationally considerably more expensive. Lesser computational complexity makes audit planning and analysis using simulations as in Section 6 more feasible.

Notice also that PROVIDENCE and MINERVA are identical for $j = 1$.

3.2 Risk-Limiting Property: Proof

We now prove that PROVIDENCE is risk-limiting against a strong adversary using lemmas from basic algebra which are given 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 round sizes used in this audit, with corresponding cumulative tallies of ballots for the reported winner $\boldsymbol{k_j}$. For round $j = 1$, by Definitions 6 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 6 and Definition 7, there is a value $k_{min,1} = k^{p_a,p_0,\alpha,0}_{min,1,0,n_1}$ such that

$$\frac{Pr[K_1 \geq k_1 \mid H_a, n_1]}{Pr[K_1 \geq k_1 \mid H_0, n_1]} \geq \frac{Pr[K_1 \geq k_{min,1} \mid H_a, n_1]}{Pr[K_1 \geq k_{min,1} \mid H_0, n_1]} \geq \frac{1}{\alpha}.$$

For any round $j \geq 2$, by Definition 6 and Lemma 6, $\mathcal{A} =$ Correct (the audit stops) if and only if

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

By Lemma 7 and Definition 3, this is equivalent to

$$\frac{\Pr[K_{j-1} = k_{j-1} \mid H_a, n_{j-1}] Pr[K_j \geq k_j \mid k_{j-1}, H_a, n_{j-1}, n_j]}{\Pr[K_{j-1} = k_{j-1} \mid H_0, n_{j-1}] Pr[K_j \geq k_j \mid k_{j-1}, H_0, n_{j-1}, n_j]} \geq \frac{1}{\alpha}.$$

By Lemma 6 and Definition 6, we see that there exists a $k_{min,j} = k^{p_a,p_0,\alpha,k_{j-1}}_{min,j,n_{j-1},n_j} \leq k_j$ for which

$$\frac{\Pr[K_{j-1} = k_{j-1} \mid H_a, n_{j-1}] Pr[K_j \geq k_j \mid k_{j-1}, H_a, n_{j-1}, n_j]}{\Pr[K_{j-1} = k_{j-1} \mid H_0, n_{j-1}] Pr[K_j \geq k_j \mid k_{j-1}, H_0, n_{j-1}, n_j]} \geq \frac{\Pr[K_{j-1} = k_{j-1} \mid H_a, n_{j-1}] Pr[K_j \geq k_{min,j} \mid k_{j-1}, H_a, n_{j-1}, n_j]}{\Pr[K_{j-1} = k_{j-1} \mid H_0, n_{j-1}] Pr[K_j \geq k_{min,j} \mid k_{j-1}, H_0, n_{j-1}, n_j]} \geq \frac{1}{\alpha}$$

The above may be rewritten as

$$\sum_{k=k_{min,j}}^{n_j} Pr[(K_j, K_{j-1}) = (k, k_{j-1}) \mid H_0, n_{j-1}, n_j] \leq \alpha \sum_{k=k_{min,j}}^{n_j} Pr[(K_j, K_{j-1}) = (k, k_{j-1}) \mid H_a, n_{j-1}, n_j]$$

The left hand side above is the probability of stopping in the j^{th} round and $K_{j-1} = k_{j-1}$, given the null hypothesis, which is smaller than α times the same probability given the alternate hypothesis. For different possible values of k_{j-1}, different round sizes n_j can be used, and this same relationship will hold. That is, the relationship holds even if the values of n_j depend on k_{j-1}, if n_j is a function $n_j(\alpha, p_0, p_1, \boldsymbol{k}_{j-1}, \boldsymbol{n}_{j-1})$.[1]

Summing both sides over all values of $k_{j-1} < k_{min,j-1}$ gives us a similar relationship between the probabilities of stopping in round j (given the null and alternate hypotheses respectively). When both sides of the inequality are further summed over all rounds, we get:

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

Finally, because the total probability of stopping the audit under the alternative hypothesis is not greater than 1, we get

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

□

3.3 Consequences of resistance to an adversary choosing round size

To illustrate the practical implication of this property, we consider a toy example: an RLA of a two-candidate contest with margin 0.01 and risk limit 0.1. Suppose we wish to achieve a conditional stopping probability 0.9 in each round of the audit. For PROVIDENCE, we can compute a new round size for each round based on the previous samples. For MINERVA,

[1] MINERVA enforces a similar relationship between risk and stopping probability but does so at the level of the round rather for each individual value of K_{j-1}. By enforcing this relationship for each value of K_{j-1}, PROVIDENCE is resistant to a strong adversary.

however, we would have a predetermined round schedule. We use the default MINERVA round schedule of audit software Arlo [21] (used by many states performing an RLA), which is $[x, 2.5x, 6.25x, ...]$; that is, the next marginal round size is 1.5 times the current one. This multiplier of 1.5 is known to give, over a wide range of margins, a probability of stopping roughly 0.9 in the second round if the first round size has probability of stopping 0.9.

Both the audits of our toy example therefore begin with a first round size of 17,272 with a 0.9 probability of stopping, and both will stop in the first round if the sample contains at least 8,725 ballots for the winner. We now consider two cases for which the audit proceeds to a second round.

In one case there are 8,724 votes for the winner in the sample, just one fewer than the minimum needed to meet the risk limit. In the MINERVA audit, we are already committed to a second round size of 43,180 which, given the nearly-passing sample of the first round is higher than necessary, achieving a stopping probability in the second round of .954. The PROVIDENCE audit samples more than 9,000 fewer ballots with a round size of 34,078, achieving the desired 0.9 probability of stopping.

In a less lucky sample, the winner recieves 8,637 ballots, few more than the loser recieves. In the MINERVA audit, we again have to use a second round size of 43,180, but now this round size only achieves a 0.727 probability of stopping, significantly less than the desired 0.9. Again, the PROVIDENCE audit can scale up the second round size according to the first sample and achieve the desired 0.9 probability of stopping with 58,007 ballots.

3.4 Efficiency

Lemma 1. *For any risk-limit $\alpha \in (0,1)$, for any margin and for any round schedule $[n_1, \ldots, n_j]$, the* PROVIDENCE *RLA stops before or in the same round as EoR* BRAVO.

Proof. Appendix A □

4 PROVIDENCE Audit Simulations

We use simulations to provide additional evidence for our theoretical claims regarding PROVIDENCE and to gain insight into audit behavior. As done in [6], we use margins from the 2020 US Presidential election—state-wide pairwise margins of 0.05 or larger between the two leading candidates. Narrower margins are computationally expensive, especially for the simulations of tied elections, which, by design, have a low probability of stopping and hence quickly increase in sample size. We use the simulator in the R2B2 software library [1]. For each margin, we perform 10^4 PROVIDENCE audit trials each on a tied election (hypothesis H_0, the null hypothesis) and the election as reported (hypothesis H_a, the alternate hypothesis). All trials have risk limit $\alpha = 0.1$, a maximum of 5 rounds, and a conditional stopping probability of 0.90 in each round. That is, each next round size is selected to be large enough to give a 0.90 conditional probability of stopping in that round, assuming the announced tally is correct and given the tally of previous rounds. We use a maximum of five rounds because virtually no audits would progress beyond five rounds given the large conditional probability of stopping.

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.

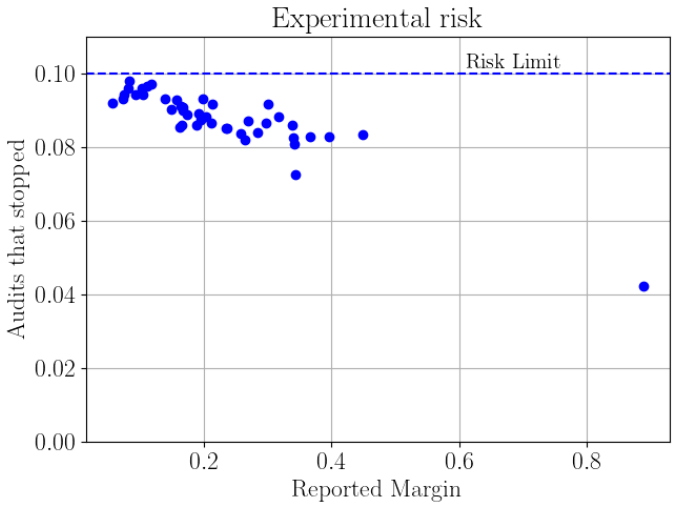

Figure 1: The fraction of simulated PROVIDENCE audits on tied elections that stopped in any rounds (we performed five rounds at a risk limit of 0.1) as a function of contest margin. This value is an estimate of the maximum risk of the PROVIDENCE audit.

Simulations of audits of the election as reported provide insight into stopping probability and number of ballots drawn when the election is as reported. Figure 2 shows that the stopping probabilities over the first rounds are near and slightly above 0.9 as expected, since our software chose round sizes to give at least a 0.9 conditional stopping probability. The values are not as tight around 0.9 for later rounds because fewer audit trials make it to later rounds, and our experimental probability estimates are not as accurate.

We now investigate the efficiency of PROVIDENCE compared to MINERVA, SO BRAVO, and EoR BRAVO by taking a single margin as an example: the 2020 US Presidential election in the state of Texas, with margin 0.057. We run an additional 10^4 simulations for each of the three other audits on the same underlying election and on a tied election. Both BRAVO implementations use a conditional stopping probabil-

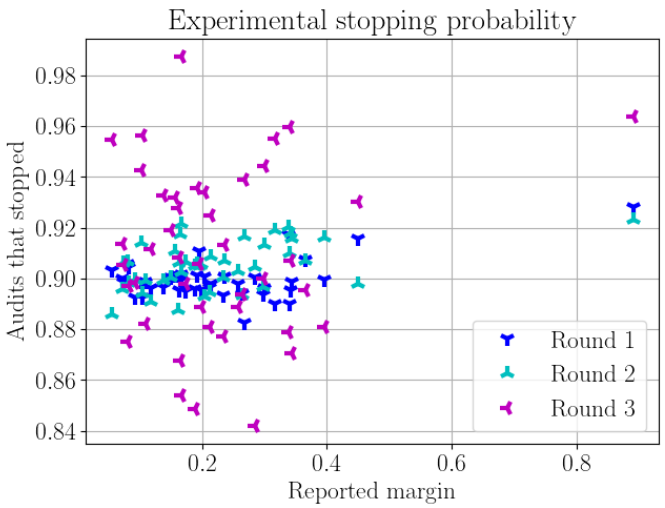

Figure 2: The fraction of simulated PROVIDENCE audits of the election as reported that stopped for each round as a function of margin. This value is an estimate of the stopping probability conditioned on the sample of the previous round. The average fraction for rounds 1, 2, and 3 is 0.8996, 0.9052, and 0.9098 respectively. We show only the first three rounds since so few audits make it to rounds 4 and 5 (of the order of $10^4 \times (0.1)^3$ and $10^4 \times (0.1)^4$ respectively).

ity of 0.9 for each round, while MINERVA uses a first round size with stopping probability 0.9 and a multiplier of 1.5 to obtain subsequent round sizes.

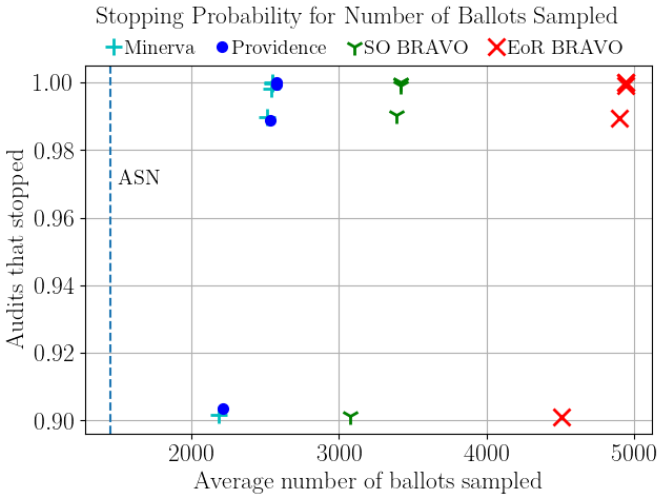

Figure 3: For the entire audit, consisting of all five rounds, the fraction of simulated audits that stopped as a function of the average number of ballots drawn for PROVIDENCE, MINERVA, EoR BRAVO, and SO BRAVO. The average sample number (ASN) for B2 BRAVO is included for context.

Figure 3 shows the probability of stopping as a function of the number of ballots sampled, a plot similar to those presented in [6]. 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 smaller than 0.03) the difference in number of ballots sampled could correspond to many days of work which would need to be completed before a certification deadline.

5 Pilot use

The Rhode Island Board of Elections performed a pilot audit in the city of Providence in February 2022. The contest audited was a single yes-or-no question in the November 2021 election: Portsmouth's Issue 1, "School Construction and Renovation Projects". The question had a reported margin of 0.2567 and the audit used a risk-limit of 0.10.

A first round size of 140 ballots with large probability of stopping (0.95) was selected. Selection order was tracked for the sake of analysis. As expected, the audit concluded in the first round. The PROVIDENCE risk was 0.0418. Table 1 shows risk measures for the drawn sample using PROVIDENCE, MINERVA and BRAVO (both EoR and SO).

ballots	PROVIDENCE	MINERVA	SO BRAVO	EoR BRAVO
140	**0.0418**	**0.0418**	**0.0541**	0.366

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.

Note that the risk measures shown in Table 1 imply that, for the sample obtained in the pilot audit, an EoR BRAVO audit would not have stopped in the first round, despite the large round size. Further, if the risk limit had been 0.05 instead of 0.10, SO BRAVO also would have required moving on to a second round.

We can use simulations to better understand typical audit behavior for the margin of this pilot audit and contextualize the results we obtained in the pilot. We run 10^4 trial audits for several stopping probabilities p. Each round size is chosen to give a probability of stopping p assuming the announced tally and given the results of previous rounds. We use the same 0.1 risk limit and margin of 0.2557.

Figure 4 shows the average number of ballots sampled for each value of p in the simulations. The vertical line denotes the stopping probability of the first round size actually chosen in the pilot (140 ballots). The large value of p corresponds

to a large first round size and a corresponding large value of average number of ballots. In later sections we show why average number of ballots is not the only metric to optimize, and how large round sizes can be beneficial from the perspective of other important metrics.

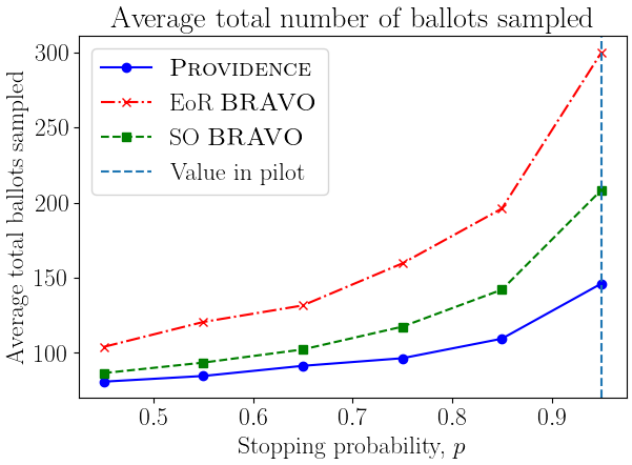

Figure 4: The total number of ballots sampled on average as a function of p, the conditional stopping probability used to select each round size. We use the same contest parameters and risk limit as the Rhode Island pilot.

For this pilot audit, extensive planning of the round schedule was not necessary because the margin was large enough that relatively few ballots were needed to achieve the high probability of stopping. In Section 6 we consider a larger statewide contest in Virginia, where selecting the round schedule has more significant implications. Virginia also currently uses ballot polling RLAs, whereas Rhode Island primarily uses batch comparison RLAs. Some of the ideas introduced in Section 6 provide a context for this pilot case as well.

For the sake of analysis, the selection order of the ballots sampled during the pilot was also recorded. Figure 5 shows the cumulative tally of winner ballots after each new ballot in the selection order is added to the sample. We observe two interesting phenomena in this particular sample's selection order.

First, an SO BRAVO audit of this sample stops because the BRAVO condition is met when the sample (orange line) surpasses the minimum number of winner ballots (blue line) earlier in the sample.[2] EoR BRAVO, however, does not stop. It might be difficult to explain to the public why SO BRAVO stops in more extreme cases like this, where the condition is met early in the sample, but poor evidence for the alternative hypothesis in the rest of the sample is ignored.

Second, only 5 of the first 11 ballots were for the announced winner. A first round of size 11 would have resulted in a smaller average total number of ballots drawn, but would have provided a misleading sample (suggesting that the winner was incorrectly reported) due to a too-small sample size.

[2]Such cases also provide insight into how PROVIDENCE is a tighter test in expectation because SO BRAVO ignores information from the rest of the sample after the BRAVO condition is met at some point earlier in the selection order.

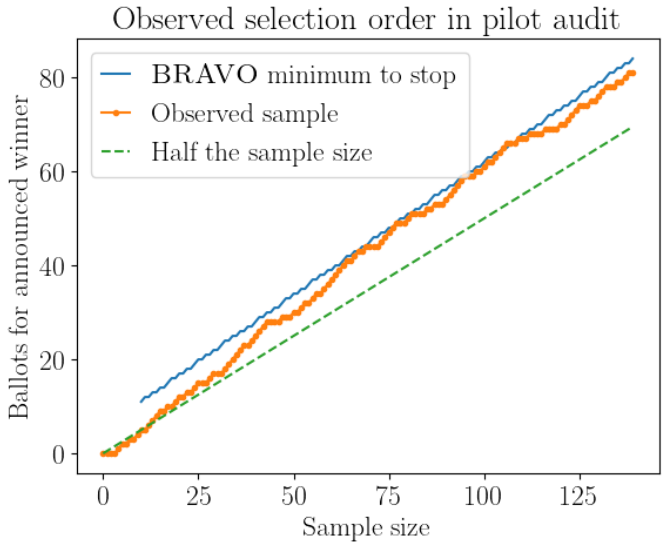

Figure 5: For each sample size from 1 to 140, the intermediate cumulative sum of ballots for the announced winner found in the sample is shown.

Both these ideas are addressed more thoroughly in Section 6.

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 are not large enough to accurately capture the distribution of votes. There is a higher probability that 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. Using 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 0.053 between the two candidates with the most votes. Analytical approximation of the expected audit behavior (quantities like expected total number of ballots sampled or total number of rounds) 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 0.1.

We simulate audits considering each candidate with a column in the results available at the Virginia Department of Elections website, including irrelevant ballots. 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. Note that since there are multiple candidates, we compute the minimum round size to achieve stopping probability p for each pairwise contest between the winner and one of the losers, and we then select the largest such minimum round size and scale it up according to the proportion of the total ballots that are relevant to that pairwise contest. For this round schedule scheme, a one-and-done audit is achieved by choosing large p, say $p = .9$ or $p = .95$. We run 10^4 trial audits for each value of p, assuming the reported results are correct[3].

Note that simulations of audits of tied elections are not necessary, as all the audits we are considering are risk-limiting and hence we already know the performance to expect when auditing a tied election, even one not reported as such.

Importantly, note that MINERVA does not appear in the analysis in this section. Questions about the efficiency of MINERVA for its necessarily fixed round schedules are addressed in Section 4, but in this section round sizes are chosen to have specified probabilities of stopping given previous samples. MINERVA is not known to be risk-limiting in this setting, and thus cannot be used for RLAs that proceed in this way.

6.1 Person-hours

6.1.1 Average total ballots.

The simplest workload models are a function of just the total number of a ballots sampled.[4] Figure 6 shows the average total number of ballots sampled as a function of p.

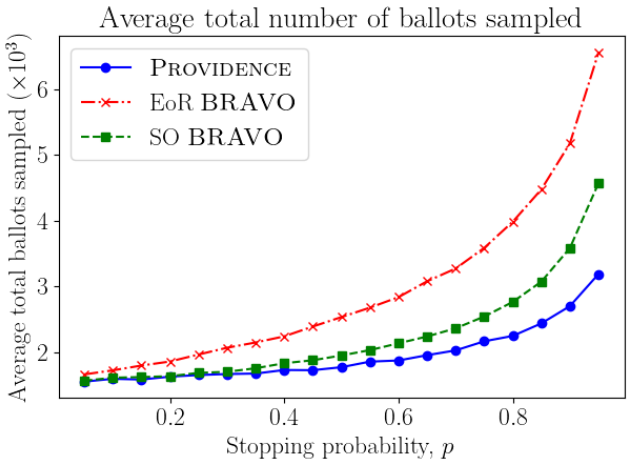

Figure 6: The total number of ballots sampled on average, as a function of p, the conditional stopping probability used to select each round size, for ballot polling audits of the 2020 US Presidential election in the state of Virginia.

Figure 7 provides the same number as a fraction of the PROVIDENCE values. It is straightforward to show that PROVIDENCE and both forms of BRAVO collapse to the same test when each round corresponds to a single ballot. Figures 6 and 7 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 $p = 0.95$, EoR BRAVO and SO BRAVO require more than 2 and 1.4 times as many ballots as PROVIDENCE respectively.

6.1.2 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 after drawing them from secure storage boxes stored in county locations. 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

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

[4] Sometimes total *distinct* ballots sampled is used, but for the margins we use in our examples in this section, the difference between total distinct ballots and total ballots is insignificant [23]. It is straightforward to modify the model we discuss here to account for total distinct ballots.

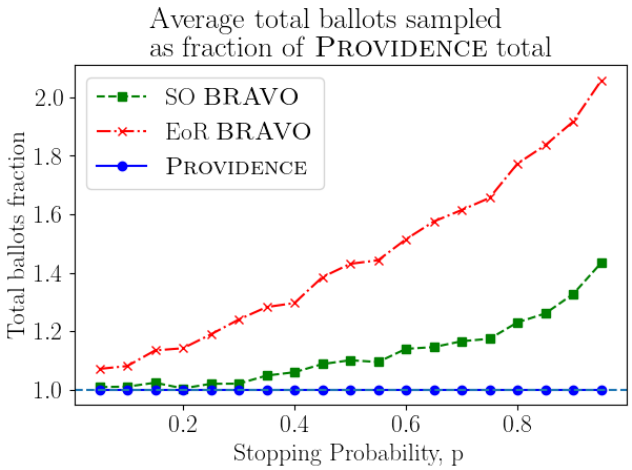

Figure 7: The total number of ballots sampled on average, as a fraction of those sampled by PROVIDENCE, as a function of p, the conditional stopping probability used to select each round size, for ballot polling audits of the 2020 US Presidential election in the state of Virginia.

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 to represent, for example, the effort of constructing a ballot manifest. For simplicity, (and without loss of generality), we measure in multiples of the per ballot workload; that is, we assume it is one unit, $w_b = 1$. A per round workload of $w_r = x$ corresponds to a per round workload which is x times the per ballot workload. We use $w_r = 1000$ 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 [7], 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, on average, per round in each of the 133 counties of Virginia, a clearly conservative workload estimate.

As shown in Figure 8, average workloads first reduce as stopping probability increases; this is likely due to a decrease in the number of rounds. After hitting a sweet spot, average workloads again increase with stopping probability; this time, likely because the average number of rounds does not increase much and the cost changes because of number of ballots drawn, which increases with round size. PROVIDENCE achieves the lowest minimum average workload at roughly $p = 0.7$ for our example choice of $w_r = 1000$.

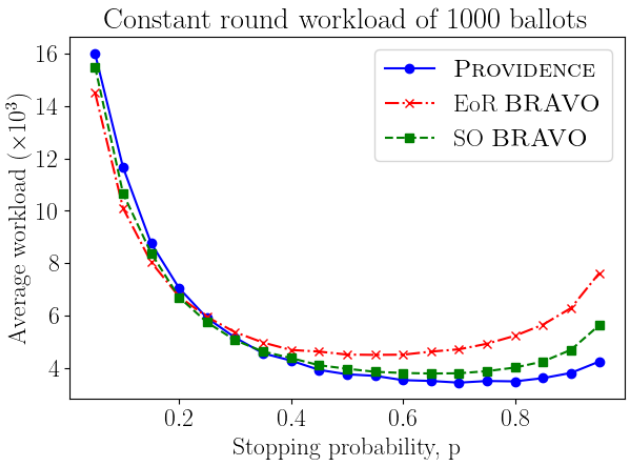

Figure 8: For workload parameters $w_b = 1$ and $w_r = 1000$, this plot shows the expected workload for various values of p. Expected workload is found using Equation 4 and the average number of ballots and rounds in our simulations as the expected number of ballots and rounds.

Importantly, this gives us a way to estimate the minimum 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 8 and then find the minimum average workload achieved for each of the audits and its corresponding stopping probability p.

Figure 9 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 workload is minimized by minimizing the number of ballots drawn, which corresponds to small round sizes, and we would expect all three audits to behave similarly, as ballot-by-ballot audits, with the smallest workload. On the other hand, for extremely large 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, all corresponding to large round sizes in order to minimize the number of rounds. We know that there is variation in the number of ballots used by each type of audit for large round sizes (a factor of two for $p = 0.9$), but these values would be small in comparison to w_r. We observe this behaviour in Figure 9 for extremely small and large workload 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 10 shows the corresponding round schedule parameters p that achieve these minimal workloads. As expected,

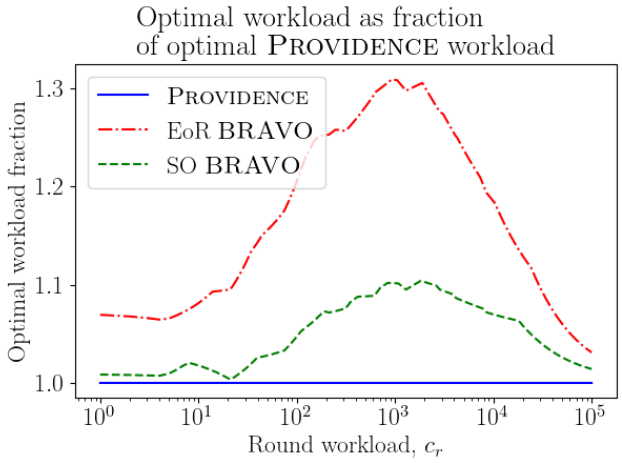

Figure 9: For varying round workload w_r, the optimal average workload achievable by each audit, as a fraction of the PROVIDENCE values.

an overhead for each round means that larger round sizes are needed to achieve an optimal audit, and so for all three audits 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.

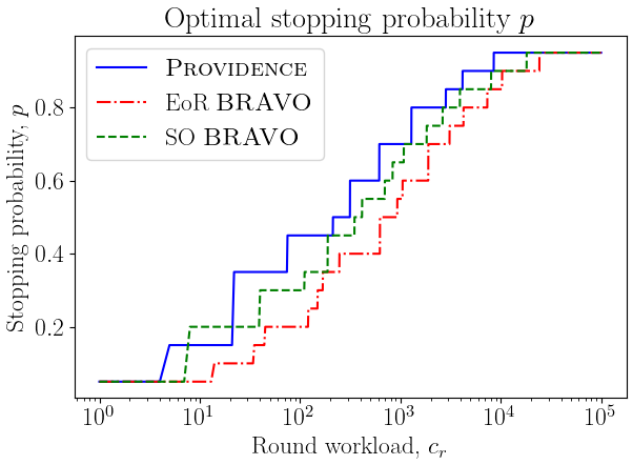

Figure 10: The optimal (workload-minimizing) stopping probability p for varying workload model parameters w_r. (Note that the steps in this function are a consequence of our subsampling the workload function. That is, the workload-minimizing value of p for each w_r is only allowed to take on values at increments of 0.05.)

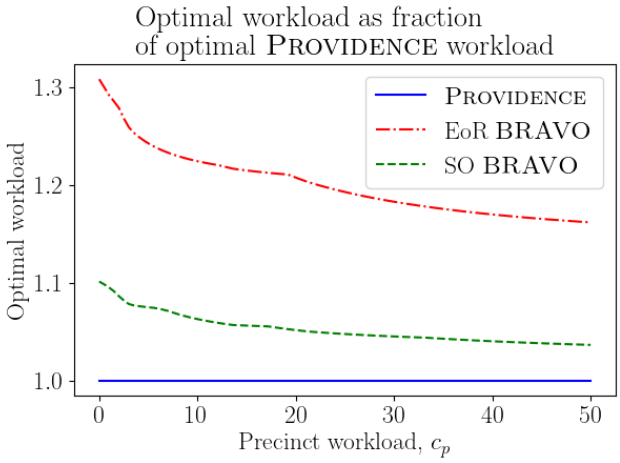

Figure 11: Optimal average workload using the workload Equation 5 for varying w_p, given as a fraction of the value for PROVIDENCE. Similar to Figure 9, we show a generous range of values for the workload variable, c_p in this case. If the time for a single ballot is 75 seconds, then $c_p = 50$ corresponds to over an hour of extra time to sample a ballot from a new container.

6.1.3 Precinct overhead.

For a more complete model, we can also introduce container-level workload. If a round requires multiple ballots from a single container, the container need only be unsealed once. Based on a Rhode Island pilot RLA report [7], 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 as a per-precinct 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 11.

6.2 Real time

Given tight certification deadlines[5], the total real time to conduct the RLA is also an important factor to consider when

[5]Virginia recently passed legislation requiring pre-certification RLAs.

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. Note that some other county may be slower, having fewer votes but also less auditing resources; but still, a slowest county exists. In this example, we take it to be Fairfax, the largest. For the slowest 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 $\bar{E}_p$. 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

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

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.[6] In practice, election officials could use our software and their own estimates of these values to explore choices for round schedules. Figure 12 shows how the estimated 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 retrieving and checking ballots, the distribution of ballots and containers both in number and physical space, and other factors. We provide Figure 12 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 have a legitimate need to include personal safety as a consideration. In a random sample, a true loser may receive more votes than the true winner. This happens more often when the sample sizes are small, like for a hypothetical first round size of 11 in the pilot audit, as seen in Figure 5. In the abstract, 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.

[6]The value $t_b = 75$ seconds corresponds to a serial retrieval and interpretation of the ballots based on the [7] timing, $t_p = 75$ seconds corresponds to the approximate doubling in time for new-box ballots as reported in [7] in the ballot-level comparison timing data, and $t_r = 3$ hours is just a guess at an approximate order for this variable.

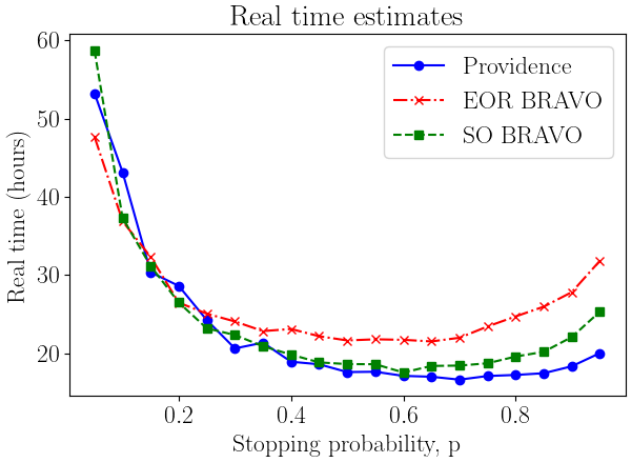

Figure 12: The real time as estimated by Equation 6 for varying p with expected values as estimated by our simulations.

Imagine that Alice beats Bob in an election contest both truly and in the reported 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 (and drawing fewer ballots on average). 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, as it can appear to Alice's supporters that election officials are simply drawing ballots till a chosen outcome is obtained.

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 13 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 13, 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 needed, and to achieve a probability of misleading of roughly

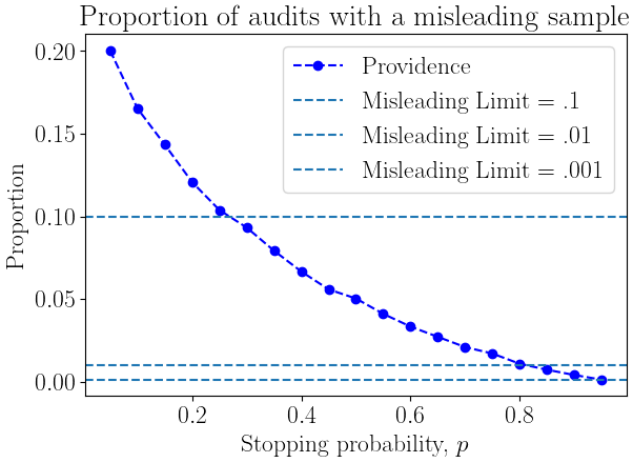

Figure 13: The proportion of simulated PROVIDENCE audits for the Virginia contest parameters that had a *misleading sample* in any round.

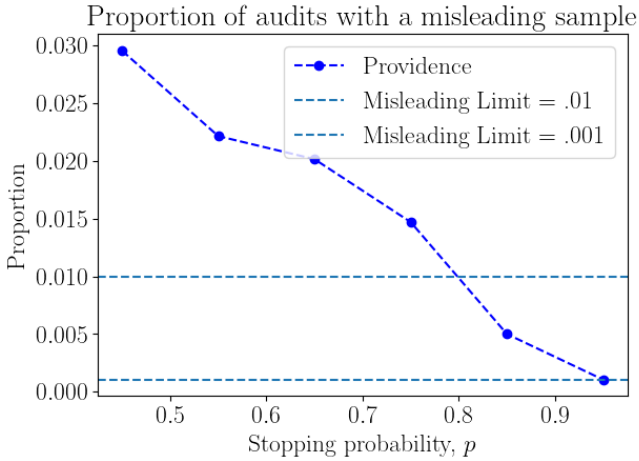

Figure 14: The proportion of simulated PROVIDENCE audits for the pilot audit parameters that had a *misleading sample* in any round.

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.

We observe a similar behavior in our simulations of audits on the contest from the pilot audit. Figure 14 shows the proportion of the pilot simulations which contained a *misleading sample* in any round. Despite the large difference in margin (~ 0.05 in Virginia and ~ 0.25 in the pilot) we still observe that a *misleading limit* of 0.01 is first achieved at roughly $p = 0.8$ and 0.001 at $p = 0.95$.

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.

6.3.1 Misleading SO BRAVO sequences.

As we consider the idea of misleading samples, it is noteworthy that SO BRAVO suffers from a different and unique type of misleading result.

After drawing a cumulative $n > 1$ 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 a value of k, however, the particular sequence of the sample that led to that value of k contains no additional information about whether the sample is more likely under the alternative or null hypotheses. That is to say, $\Pr[K = k|H_a]$ and $\Pr[K = k|H_0]$ have the same value regardless of the sequence. Despite this, the SO BRAVO RLA stopping condition is not just a function of n and k but also a function of the sequence, the selection order. 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. We refer to such sequences which stop under SO BRAVO, but not under EoR BRAVO, as *misleading sequences*. To be clear, this is not a mathematical issue; stopping in such cases is still a correct application of Wald's SPRT result [22]. The misleading nature of such stoppages is the note we are making. This is another case where election officials might have difficulty explaining the misleading situation to the public.

Recall from Section 5 that the pilot PROVIDENCE RLA performed in Providence, Rhode Island had an SO BRAVO *misleading sequence*. In particular, the audit passed with an SO BRAVO risk measure of 0.0541 but the final cumulative tally of the sample gives a BRAVO risk measure of 0.366.

It is easy to use our simulations to see how often SO BRAVO

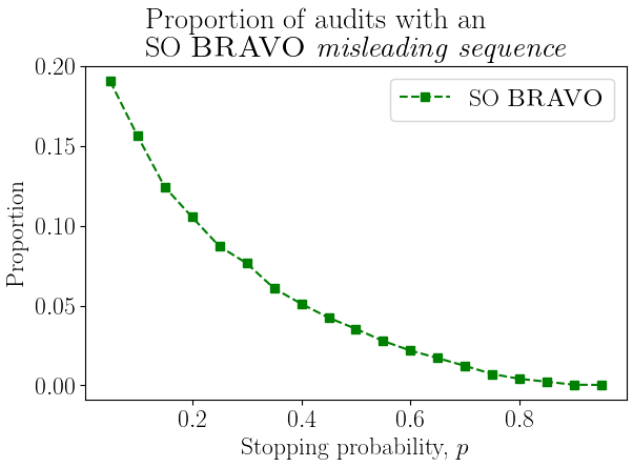

Figure 15: The proportion of sequences that are misleading sequences in the SO BRAVO audit as a function of p.

misleading sequences occur by checking whether the final cumulative sample of each SO BRAVO trial meets the EoR BRAVO stopping condition and counting those which do not. Figure 15 shows the proportion of simulated SO BRAVO audits that stopped with a *misleading sequence*. Unlike the more general *misleading sample* discussed so far, these *misleading sequences* are unique to SO BRAVO audits, and Figure 15 only shows the proportion of audits that stopped with a *misleading sequence*; additional SO BRAVO audits also contained *misleading samples*.

7 Conclusion

A rigorous tabulation audit is an important part of a secure election. We present PROVIDENCE and demonstrate that it is as efficient as MINERVA and as 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 for planning audits. We define the constraint of an acceptable probability of a misleading audit sample, and describe its importance to the planning process.

8 Availability

PROVIDENCE is implemented in the open source R2B2 software library for R2 and B2 audits [1] We provide software to test stopping conditions, find round sizes to achieve a given probability of stopping, and find round sizes that have acceptable probabilities of a *misleading sample*. The software also includes the simulator for all these audits and the functionality to perform the workload and real time analysis we present in this paper.

M	margin	n	Prov	SO	EoR
0.1	0.25	25	0.221	0.152	0.115
	0.15	73	0.202	0.186	0.141
	0.05	657	0.227	0.192	0.127
	0.03	1825	0.246	0.194	0.124
	0.01	16423	0.246	0.196	0.124
0.01	0.25	85	0.792	0.707	0.559
	0.15	239	0.817	0.712	0.549
	0.05	2163	0.817	0.721	0.569
	0.03	6011	0.824	0.723	0.573
	0.01	54117	0.824	0.724	0.57
0.001	0.25	149	0.962	0.889	0.783
	0.15	421	0.958	0.894	0.801
	0.05	3815	0.96	0.896	0.785
	0.03	10607	0.961	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.

9 Acknowledgements

The authors are grateful to the Rhode Island Board of Elections for conducting a pilot PROVIDENCE RLA, and to Georgina Cannan, Liz Howard, Mark Lindeman, and John Marion for their support of the pilot. The authors thank Audrey Malagon for useful information on audits.

References

[1] Anonymized. The R2B2 (Round-by-Round, Ballot-by-Ballot) library, https://github.com/xxxx.

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

[3] Matthew Bernhard. Risk-limiting Audits: A practical systematization of knowledge. In *In Proceedings, Seventh International Joint Conference on Electronic Voting (E-Vote-ID'21), October 2021*, 2021.

[4] 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.

[5] Jennie Bretschneider, Sean Flaherty, Susannah Goodman, Mark Halvorson, Roger Johnston, Mark Lindeman, Ronald L. Rivest, Pam Smith, and Philip B. Stark. Risk-limiting post-election audits: Risk-limiting post-election audits: Why and how. https://www.stat.berkeley.edu/ stark/Preprints/RLAwhitepaper12.pdf, October 2012.

[6] 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.

[7] 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.

[8] Lynn Garland, Mark Lindeman, Neal McBurnett, Jennifer Morrell, Marian Schneider, and Stephanie Singer. Principles and best practices for principles and best practices for post-election tabulation audits. https://verifiedvoting.org/wp-content/uploads/2020/05/Principles-and-Best-Practices-For-Post-Election-Tabulation-Audits.pdf, December 2018.

[9] 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.

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

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

[12] 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.

[13] Katherine McLaughlin and Philip B. Stark. Workload estimates for risk-limiting audits of large contests. Honors Thesis, University of California, Berkeley, 2011.

[14] 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.

[15] 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.

[16] 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.

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

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

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

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

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

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

[23] 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.

[24] 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 any risk-limit $\alpha \in (0,1)$, for any margin and for any round schedule $[n_1,\ldots,n_j]$, the PROVIDENCE RLA is more efficient than EoR BRAVO.

Proof. Let $[n_1,\ldots,n_j]$ be a round schedule, and assume that an EoR BRAVO audit stops in round j, after observing $k_1,\ldots,k_j$ ballots for the announced winner in each round

respectively. That is, the EoR BRAVO stopping condition is true:

$$\sigma(k_j, p_a, p_0, n_j) \geq \frac{1}{\alpha}.$$

To see the PROVIDENCE stopping condition is fulfilled, we rewrite as

$$\frac{1}{\alpha} \leq \sigma(k_j, p_a, p_0, n_j)$$

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

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

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

Where inequality $(*)$ follows from [23, Theorem 6]. Note that we apply this result on τ_j for just $j = 1$.

□

Lemma 2. *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 $0 \leq k \leq n$.*

Proof. See [24, Lemma 4]. □

Lemma 3. *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. See [24, Lemma 2]. □

Lemma 4. *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 of k for $0 \leq k \leq n$.*

Proof. Apply Lemmas 2-3. □

Lemma 5. *Given a strictly monotone increasing sequence: $x_1, x_2, \ldots x_n$ and some constant A,*

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

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

Proof. Evident. □

Lemma 6. *For $\mathcal{A} = (\alpha, p_a, p_0, k_{j-1}, n_{j-1}, n_j)$-PROVIDENCE, there exists a $k_{min,j,n_{j-1},n_j}^{p_a,p_0,\alpha,k_{j-1}} = k_{min,j}(\text{PROVIDENCE}, p_a, p_0, k_{j-1}, n_{j-1}, n_j)$ such that*

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

Proof. From Definition 6,

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

Now to apply Lemma 5, 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 4. For $j \geq 2$,

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

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

As a function of k_j, σ is constant, and thus ω is strictly increasing by Lemma 4. Therefore by Lemma 5, we have the desired property. □

Lemma 7. *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

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

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

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

□

Definition 7. *Let $[n_1, \ldots, n_j]$ be the round schedule of an audit that has not stopped by the round $j-1$. Let us define*

$$k_{min,j,n_{j-1},n_j}^{p_a,p_0,\alpha,k_{j-1}} = \min\left\{k : \omega_j(k, k_{j-1}, p_a, p_0, n_j, n_{j-1}) \geq \frac{1}{\alpha}\right\}. \quad (7)$$

As we have seen in Lemma 6, such a value of $k_{min,j,n_{j-1},n_j}^{p_a,p_0,\alpha,k_{j-1}}$ exists and $k_j \geq k_{min,j,n_{j-1},n_j}^{p_a,p_0,\alpha,k_{j-1}}$ if and only if the result of the audit is Correct, (*i.e.,* the stopping condition in Definition 6 holds).

The following lemma shows a Markov-like property of PROVIDENCE audit (*i.e.,* for an audit that has not stopped in the first $j-1$ rounds, only cumulative results of the round $j-1$ matter: cumulative sample size n_{j-1} and the number of ballots for the winner k_{j-1}).

Lemma 8. *Let $[n_1,\ldots,n_{j-1},n_j]$ be a round schedule for an execution of* PROVIDENCE *audit that has not stopped in any of its first $j-1$ rounds (i.e., for every $i=1,\ldots,j-1$: $k_i < k_{min,j,n_{j-1},n_j}^{p_a,p_0,\alpha,k_{j-1}}$), then:*

$$k_{min,j,n_{j-1},n_j}^{p_a,p_0,\alpha,k_{j-1}} = k_{min,2,n_{j-1},n_j}^{p_a,p_0,\alpha,k_{j-1}}.$$

Proof. Let k_{j-1} denote the number of ballots drawn for the declared winner up to the round $j-1$ (out of n_{j-1} sampled ballots). The stopping decision for the round j is made as follows:

$$k_{min,j,n_{j-1},n_j}^{p_a,p_0,\alpha,k_{j-1}} = \min\left\{k : \omega_j(k,k_{r-1},p_a,p_0,n_r,n_{r-1}) \geq \frac{1}{\alpha}\right\} =$$

$$= k_{min,2,n_{j-1},n_j}^{p_a,p_0,\alpha,k_{j-1}}$$

□

That is, the stopping condition is equivalent to that of a two round audit with the same cumulative votes for the winner and cumulative round sizes: the first round is of size n_{j-1} and has k_{j-1} votes for the winner, and the second (cumulative) round size is n_j with k_j (cumulative) votes for the winner. Compare this to the similar property for the BRAVO stopping condition.