Risk-Limiting Audit PROVIDENCE and Round Size Considerations

by Oliver Broadrick

B.S. in Computer Science, May 2022, The George Washington University

A Thesis submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial satisfaction of the requirements
for the degree of Master of Science

May 21, 2023

Thesis directed by

Poorvi L. Vora
Professor of Computer Science

Filip Zagórski
Assistant Professor of Computer Science, University of Wrocław

Dedication

It’s all Poorvi’s fault.

Acknowledgments

We are grateful to the Rhode Island Board of Elections for conducting a pilot PROVIDENCE RLA. The audit is named in recognition of their efforts. We thank Georgina Cannan, Liz Howard, Mark Lindeman, and John Marion for their support of the pilot; Audrey Malagon for useful information on audits; and an anonymous *USENIX Security 2023* shepherd for very useful guidance on the presentation of our work. Conversations with Philip B. Stark provided much insight.

Abstract

Risk-Limiting Audit PROVIDENCE and Round Size Considerations

A Risk-Limiting Audit (RLA) is a statistical election tabulation audit with a rigorous error guarantee: viewed as a binary hypothesis test with the null hypothesis being that the announced election outcome is incorrect, its Type I error is bounded above, whatever the true election tally. We present ballot polling RLA PROVIDENCE[1], an audit with the efficiency of RLA MINERVA and flexibility of RLA BRAVO, and prove that it 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 sample when rounds are too small, demonstrating the importance of the resulting constraint on audit planning. We describe an approach to planning audit round schedules using these measures and present simulation results demonstrating the superiority of PROVIDENCE.

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 in the United States.

[1] A shorter version of this work [6] will appear at USENIX Security 2023.

Table of Contents

List of Figures

List of Tables

Chapter 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 want to know when they do. *Software independent* voting systems [19, 18] are ones where an undetected change in the software cannot lead to an undetectable change in the election outcome. The use of software-independent voting systems is, however, not sufficient to ensure election integrity. The *evidence-based elections* [22] approach goes further and requires that election outcomes be supported by strong affirmative evidence. The evidence is generated during the election and publicly examined at the end, enabling citizens to determine whether it provides strong support of the election outcome or is not sufficiently convincing. 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 procedures were followed; and a rigorous tabulation audit, known as a risk-limiting audit (RLA) [11], to determine whether the outcome is correctly computed from the stored ballots.

As a formalized approach to the examination of evidence supporting vote tabulation, an RLA is an important part of an evidence-based election. When correctly implemented, it serves to vastly improve the trustworthiness of the election.

Significant efforts by nationwide organizations (such as Verified Voting, Brennan Center for Justice, Common Cause and Democracy Fund), local organizations, citizen advocates and experts have led to great progress towards the use of RLAs. Nonprofit VotingWorks has developed open-source election audit software, Arlo [25], and provides training in its use. Six states (Colorado, Georgia, Nevada, Pennsylvania, Rhode Island, Virginia) now require RLAs; three have statutory pilot programs (Indiana, Kentucky, Texas); four allow RLAs to satisfy a more general audit requirement (California, Ohio, Oregon, Washington);

and two have an administrative pilot program (Michigan, New Jersey) [24]. The effort to broaden the reach of RLAs continues, as election officials appear very keen to improve the trustworthiness of the elections they administer. RLAs are nowhere close to being routine though, and difficulties could reverse the gains that have been made.

1.1 Background on RLAs

A tabulation audit will either end with a declaration that an election outcome is correct, or escalate to a full hand count. The integrity of an audit may be judged by how it deals with incorrect outcomes. A *risk-limiting audit (RLA)* guarantees a minimum probability that it will perform as it is supposed to (escalate to a full hand count) if the outcome is incorrect. Equivalently, this is also a guarantee of a maximum probability with which it will perform erroneously (declare the audit correct) when the outcome is incorrect. The *risk limit* of an RLA is the (guaranteed) maximum probability that an incorrect election outcome would be declared correct. Lower risk limits are better.

There are three types of RLAs: ballot comparison RLAs, batch comparison RLAs and ballot polling RLAs. 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), real statewide audits (Georgia, Virginia) [24] and audits of smaller jurisdictions, such as Montgomery County, Ohio [29]. Ballot and batch comparison RLAs are described in Chapter 1.2.

1.1.1 The workflow of a ballot polling RLA

A ballot polling RLA [11] is based on manual examination of the sampled ballots, and does not require any information from the tabulating system other than the tally. More detail about the storage of ballots is required, however: 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)

to enable the sampling of specific ballots from the list.

All RLAs draw one or more ballots at a time; each such set of ballots is referred to as a *round*. We use notation and terminology from [29, 28, 5, 12] and also assume ballots are drawn with replacement.

Ballot polling audits proceed as follows.

1. The ballot manifest is published.

2. A first round size [29] is chosen.

3. Ballots on the ballot manifest are sampled uniformly at random, with replacement, using a pseudorandom number generator—typically seeded by a natural source of randomness like dice rolls.

4. The physical ballots are found and manually interpreted; the interpretations are recorded.

5. The stopping condition $\mathscr{A}$, a function of the manual interpretations of the current cumulative sample of ballots X, is computed. It outputs:

   (a) *Correct: stop the audit* or

   (b) *Undetermined: sample more ballots*.

   Election officials may choose to abort this procedure and go to a full hand count at any time and should have a plan for how to decide whether to do so; we discuss this in more detail below and in Chapter 2.

6. If more ballots are to be drawn, the next round size is chosen, and the audit goes back to step 2.

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 legislation, the successful

completion of the RLA (or a full hand count confirming the certified outcome) is usually necessary before certification[1] and certification deadlines would play a large role in round size and hand count decisions.

1.1.2 The audit model

An audit is typically defined as a binary hypothesis test. If the null, H_0, is defined as the incorrect outcome hardest to detect (generally a tie, see [23]), we have the following definition of an RLA.

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

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

Definition 1 is valid at the end of the RLA, and not at the end of each round. Thus it is not the case that an incorrect election will pass the audit if a sufficient number of rounds is drawn. In fact, the larger a sample, the more accurate the estimate of election correctness can be, and the more the audit will diverge from one that would stop (see the example below). It is possible to declare an incorrect outcome as correct even after drawing a large sample, but a good audit is less likely to make this error.

While the risk limit is an error measure, the *stopping probability* for a certain round size characterizes the efficiency of the audit if the outcome is correct.

1.1.2.1 Full hand count

It is worth noting here that if election officials do not have a plan for when they will move to a full hand count, and the election outcome is incorrect, at least a fraction $(1-\alpha)$ of the audits will *never* stop, and new rounds will continue to be drawn. See Figure 4.1

[1] If an RLA were performed after certification and determined that the outcome was incorrect, there may not be a legal means of changing the outcome and this could significantly impact citizen confidence. For example, till recently, the state of Virginia required RLAs but they were to be performed after certification and could not be used to change an outcome. This was corrected through new legislation passed in April 2022.

in Chapter 4 for the results of 10,000 audit simulations of PROVIDENCE on tied elections (risk limit of 0.1, across election margins of 0.05 and larger in the statewide contests for US President in 2020). Observe that, for each margin studied, more than 90% of the audits do not stop after five rounds. To illustrate this idea, we present two example simulations from among those used to compute the statistics reported in Chapter 4.

A total of 11,315,056 votes were cast in the 2020 US President contest in the state of Texas [8]; candidate Trump won with 5,890,347 votes, and the second highest vote count was that of candidate Biden, who received 5,259,126 votes, for a margin of 0.0566 in the pairwise contest. A pairwise contest between two candidates treats invalid votes and those received by other candidates as irrelevant to determining which of the two won the pairwise contest. Its *margin* is the difference between their vote counts as a fraction of the sum of their votes. Ballot polling audits used to audit government elections (such as BRAVO, MINERVA and PROVIDENCE) audit a multi-candidate contest by conducting multiple audits of the pairwise contests between the winner and every other candidate.

A first round size of 2,217 corresponds to a stopping probability of 0.9 and a risk limit of $\alpha = 0.1$ for PROVIDENCE (see [29] for a description of how first round size may be computed, MINERVA and PROVIDENCE are identical in the first round). The audit $\mathscr{A}$ computes the PROVIDENCE ratio (the likelihood ratio of the hypothesis test, which is a function of the number of votes for the two candidates in the sample, see Chapter 3.3 for more detail). If the ratio is denoted ω, the audit stops when $\omega^{-1} \leq \alpha$, or equivalently (see Definition 6) if $\omega \geq \alpha^{-1}$.

In Chapter 4 we describe the results of 10,000 audit simulations for each hypothesis, assuming that the underlying true vote distribution is (a) as announced and (b) tied. We describe below a single illustrative simulation from each hypothesis.

The first simulation assumed that the tally is correct and resulted in 1,138 votes for

Round No.	Cumulative Round Size	Trump	Biden	ω^{-1}
1	2,217	1,111	1,079	0.256
2	5,970	2,940	2,953	4.251
3	16,685	8,281	8,171	3,150
4	35,096	17,320	17,264	380,220,376
5	76,979	37,943	37,868	$2.5e^{+22}$

Table 1.1: A simulation of a PROVIDENCE audit of a tied contest with an announced margin of 0.0566, and $\alpha = 0.1$. Each round size is computed for a conditional stopping probability of 0.9. The stopping condition is $\omega^{-1} \leq \alpha$.

Trump and 1,054 for Biden (the rest were irrelevant votes). The audit stopped because:

$$\omega^{-1} = 0.047 < \alpha$$

The second simulation assumed a tied election and did not stop at the end of five rounds. Table 1.1 lists the cumulative round sizes, cumulative votes for both candidates and the inverse PROVIDENCE ratio after each draw. Each round size was chosen for a conditional stopping probability of 0.9, given that the audit did not stop so far. For example, this corresponds to an approximate probability of 0.09 that the audit stops in the second round (that it does not stop in the first round and stops in the second), and so on.

To avoid a scenario where election officials fruitlessly draw round after round hoping the audit will stop when the election outcome is incorrect, it might be worthwhile for election officials to determine, before the audit begins, the latest date by which either: the audit stops, or a hand count begins, so as to be completed before certification.

1.1.2.2 The adversary in an RLA

The goal of the adversary is to increase the true risk beyond the declared risk limit—that is, given an incorrect outcome, make the audit declare it as correct with a chance larger than the risk limit. An audit is an RLA only if there is a proof of its risk-limiting property. Hence, at the very least, the adversary would need to invalidate an assumption of the proof

to obtain:

$$Pr[\mathscr{A}(X) = \text{Correct} \,|\, H_0] > \alpha$$

1.2 Related work

1.2.1 Ballot polling RLA process

Bernhard provides a good description of the RLA and its assumptions, and also describes the process on the ground [2].

Election officials typically draw ballots in large round sizes, see for example [16, 7]. Note also that, in addition to allowing users to directly enter a round size or choose the expected number of ballots drawn by BRAVO, Arlo provides choices of stopping probabilities of 0.9, 0.8 and 0.7. For the two audits we attended, election officials chose stopping probabilities of 0.9 and 0.95. Estimates of round sizes with stopping probability 0.9 for each state in the 2020 US Presidential election may be found in [29]. Thousands of ballots is quite a common estimate; many estimates are as large as tens and hundreds of thousands of ballots. 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).

1.2.2 R2 and B2 audits and the classical BRAVO audit

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. 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 [12] 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). However, BRAVO cannot be used as a B2 audit in the scenarios described in the previous paragraph.

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, and should be expected to be inefficient because it does not use the information in the ballots drawn towards the end of the round.

1.2.3 Newer ballot polling audits

The MINERVA audit [29, 28] does not need ballot order and relies only on sample and round tallies. It 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.

The ALPHA audit [20] generalizes BRAVO to gain efficiency in cases where the reported outcome is correct but the reported margin is erroneous.

1.2.4 Simulations

Ballot polling audit simulations provide a means of educating the public and election officials [21] and understanding audit properties [14, 13, 3, 10, 12]. 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 [17], a more thorough workload estimation model includes the time taken to access individual ballots[1].

Zagórski *et al.* 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 showing 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 it approaches a B2 audit, for which BRAVO is known to be most efficient.

1.2.5 Ballot and batch comparison audits

In a ballot comparison RLA [11], 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. Not all voting systems record CVRs and their use can present privacy challenges. 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. Batch comparison typically requires the sampling of a very large number of ballots, larger than polling audits except for small enough margins.

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

1.3 Gaps in prior work and our contributions

We describe relevant gaps in current knowledge of audits and then describe our contributions.

1.3.1 Limitations of MINERVA

Zagórski *et al.* prove that MINERVA [29] is risk-limiting when the number of relevant ballots drawn in each round is pre-determined before any ballots are examined. They do not address the case of a stronger 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. An open question about MINERVA is whether its RLA proof holds in this case. Can the audit insider 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 adversary who can determine next round size after examining the current sample, 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[2]. This could lead to the drawing of too few or too many ballots and greatly constrains audit planning.

The risk limit for B2, EoR and SO BRAVO is fixed independent of whether next round sizes are determined with or without knowledge of the current sample. 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.

1.3.2 Limitations in existing workload measures

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 [14, 1, 7] 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 in 1.2, election officials, on the other hand, greatly prefer drawing many ballots at once. From conversations with election officials and staff members

[2]Note that every draw may contain 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 differences in the number of ballots average out, and that a fixed draw size is sufficient, though this is not explicitly proven in [29].

of Verified Voting, Brennan Center and Common Cause who have been training election officials to perform RLAs, we estimate that 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). Finally, in the current environment of misinformation, election officials fear a misleading audit sample (with more votes for an announced losing candidate than the winner), preferring to structure audits to reduce the chances of such samples, thus implicitly choosing larger round sizes.

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 which is necessary while planning an audit.

1.3.3 Our contributions

Our primary contribution is a new RLA, PROVIDENCE, which gives the efficiency of MINERVA and is also resistant to an adversary who can choose the next round size after knowing the current sample. BRAVO is a very flexible audit enabling audit planning; with the introduction of PROVIDENCE, similar planning is possible with greater efficiency. 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 using PROVIDENCE over BRAVO could correspond to many days of work which would need to be completed before a certification deadline.

The stopping condition for MINERVA does not take into account the sample obtained

in previous rounds. The risk limit is estimated through weighted averages across multiple rounds, assuming that round sizes do not depend on the previous sample. Attempting to simplify the proof of MINERVA's risk-limiting property, we were able to define a different audit PROVIDENCE. The RLA proof for PROVIDENCE does not make an assumption about round sizes.

We provide the following:

1. Proof that PROVIDENCE is an RLA and resistant to an adversary who can choose next round sizes after knowing the current sample.

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. The implementation of PROVIDENCE has been integrated as an option in Arlo.

PROVIDENCE may be used in any audit where sampling is with replacement and audited contests may be expressed as pairwise plurality contests: for example, it can be used in plurality and majoritarian elections.

1.3.4 Organization

Chapter 2 provides preliminaries on the BRAVO and MINERVA audits. Chapter 3 describes the PROVIDENCE audit, Chapter 4 the simulations comparing the number of ballots drawn using various ballot polling audits, and Chapter 5 the use of PROVIDENCE in an audit carried out by the Board of Elections of Rhode Island. Chapter 6 presents our workload model and describes its use for a ballot polling audit using details of the 2016 US Presidential election in the state of Virginia, while Chapter 7 introduces the notion of a misleading sample, illustrating it on the Virginia details as an example. Our conclusions, the availability of an audit implementation, and a brief description of possible future work are given in Chapter 8.

Chapter 2: Preliminaries on ballot polling RLAs

2.1 BRAVO and MINERVA

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.)

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.*

$$\sigma(k, p_a, p_0, n) \triangleq \frac{p_a^k (1-p_a)^{n-k}}{p_0^k (1-p_0)^{n-k}} \tag{2.1}$$

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) [27]. Applying the general SPRT to RLAs, there would be a third output, *Full Manual Hand Count*, in addition to *Correct* and *Undetermined*. The test requires an additional error parameter β (of audit error when the outcome is correct): the probability of requiring a (unnecessary) hand

count when the outcome is correct. The test is:

$$\mathscr{A}(X) = \begin{cases} \textit{Correct} & \sigma(k, p_a, p_0, n) \geq \frac{1-\beta}{\alpha} \\ \textit{Hand Count} & \sigma(k, p_a, p_0, n) < \frac{\beta}{1-\alpha} \\ \textit{Undetermined} & \textit{else} \end{cases} \tag{2.2}$$

An example of the above use is [11]. In the more recent literature, for example [12], 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. For a given election, σ is a function only of k, and increases with k. Any non-zero value of β would reduce the value σ is being compared with, and allow $\mathscr{A}$ to declare *Correct* for lower values of k. Thus $\beta = 0$ is the most stringent of this form of audit. The choice of going to a full hand count when $\beta = 0$ will only reduce the risk because it never results in an error (by definition, the true election outcome is one that would be obtained by a full manual hand count of the ballots) and prevents future errors.

A manual hand count presents numerous logistical challenges. The decision to move to one would be influenced by the certification deadline[1], 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. A better workload model would better inform the decision. Election officials should announce ahead of time their plans for when they would abort the audit procedure and go to a full hand count, justifying it with an eye towards completing certification requirements. This would provide a more transparent process and

[1] As mentioned earlier, audits going towards statutory or legal requirements would need to be completed by the certification deadline. Pilot audits, performed after certification, usually end after a single round or fixed number of rounds, providing a measured risk (the statistical p-value) at the end of the round, and no decision regarding a full hand count needs to be made.

protect them from political pressures.

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.*

$$\tau_j(k_j, p_a, p_0, \boldsymbol{n_j}, \alpha) \triangleq \frac{Pr[K_j \geq k_j \wedge \forall_{i<j}(\mathscr{A}(X_i) \neq Correct) \mid H_a, \boldsymbol{n_j}]}{Pr[K_j \geq k_j \wedge \forall_{i<j}(\mathscr{A}(X_i) \neq Correct) \mid H_0, \boldsymbol{n_j}]} \tag{2.3}$$

Note that τ depends only on the most recent cumulative votes for the winner, k_j and not on older values. Here, if we were to condition the ratio on the vector of cumulative winner votes $\boldsymbol{k}_{j-1}$, or, equivalently, on the most recent cumulative winner votes k_{j-1}, we would get the binomial distributions and be running a new first round each time.

The ratios for BRAVO and MINERVA are for two candidates. As mentioned when describing the example in Chapter 1.1.2, audits of multiple-candidate contests are treated as multiple audits of the pairwise contests between the announced winner and all other candidates.

2.1.1 ATHENA

For some parameters (contest, risk-limit, round sizes, etc), a MINERVA audit stops – confirming the reported outcome – for cumulative winner ballot tallies which are more probable under the null hypothesis. MINERVA is still risk-limiting in these cases. That said, one may choose in addition to the MINERVA stopping rule, to require that the likelihood ratio of BRAVO be above $\frac{1}{\delta}$ for some δ. An audit enforcing the MINERVA stopping condition and this additional bound on the likelihood ratio is called an ATHENA audit [28]. For instance, to require that the audit only stop for samples that are more likely under the alternative hypothesis, $\delta = 1$.

In particular, in the j^{th} round, for reported proportion of winner ballots p_a, vector of cumulative round sizes $\mathbf{n_j}$, vector of cumulative winner ballot tallies $\mathbf{k_j}$, and risk-limit α, an ATHENA audit stops if and only if:

1. a MINERVA audit for the same parameters stops, and
2. $\sigma(k_j, p_a, \frac{1}{2}, n_j) \geq \frac{1}{\delta}$.

While $\delta = 1$ requires that the audit stops only for samples more probable under the alternative hypothesis, smaller values of δ will be even stricter. Note that an (α, α)-ATHENA audit (i.e. $\delta = \alpha$) is an EoR BRAVO audit with risk-limit α.

2.2 Minimum winner ballots

It turns out that the stopping conditions for both BRAVO and MINERVA (and ATHENA) in any round can be expressed as a comparison test where the number of winner ballots is compared with a threshold, above which the audit stops. That is, for any margin, risk limit, round, and for either BRAVO and MINERVA, there exists some k_{min} such that the audit's stopping condition is equivalent to requiring that the number of winner ballots in the sample be greater than k_{min}. Note that a ballot polling RLA need not have this property;

the audit could stop only for even numbers of winner ballots, or some other arbitrary rule, and still abide by the risk-limiting constraint. BRAVO and MINERVA, however, do have this appealing property, as shown formally in [29].

From here forward, we will use $k_{min,j}$ to refer to the minimum number of winner ballots required to meet the stopping condition in round j (other parameters being clear from context), or simply k_{min} if the round number too is clear from context.

2.3 Intuition for MINERVA

To build intuition, we use a toy example. Consider a two candidate contest with a reported margin of 0.4. That is, the reported winner is reported to have received a proportion of 0.7 of the total ballots cast. Election officials plan to conduct a ballot polling RLA with risk limit $\alpha = 0.1$. For the first round, they will draw $n_1 = 100$ ballots. Figure 2.1 shows the probability distributions over the number of winner votes for both the alternative and null hypotheses. The BRAVO k_{min} is shown as a vertical, dashed line, and the corresponding probabilities of k_{min} under each hypothesis are highlighted with thick markers in the shape of the letter 'x'. A BRAVO audit stops (confirming the reported result) if the ratio of these two points is above a threshold, $\frac{1}{\alpha}$.

Figure 2.2 shows the same distributions but with the MINERVA k_{min}. While the BRAVO audit requires that the ratio of the two probabilities of k be above $\frac{1}{\alpha}$ in order to stop, MINERVA requires that the ratio of the tails of the distributions starting at k be above $\frac{1}{\alpha}$. Observe that the tails of the distribution starting at k_{min} correspond to the probability of drawing a number of winner ballots at least k_{min}. Equivalently, this is the probability that the audit stops in this round. We refer to the probability of stopping assuming the alternative hypothesis as the *stopping probability*. Recall that the probability of stopping under the null hypothesis is the *risk*. Thus, in the first round, MINERVA bounds by $\frac{1}{\alpha}$ the ratio of the stopping probability to the risk.

Suppose that the MINERVA audit is used, and some number $k_1 < k_{min}$ winner ballots

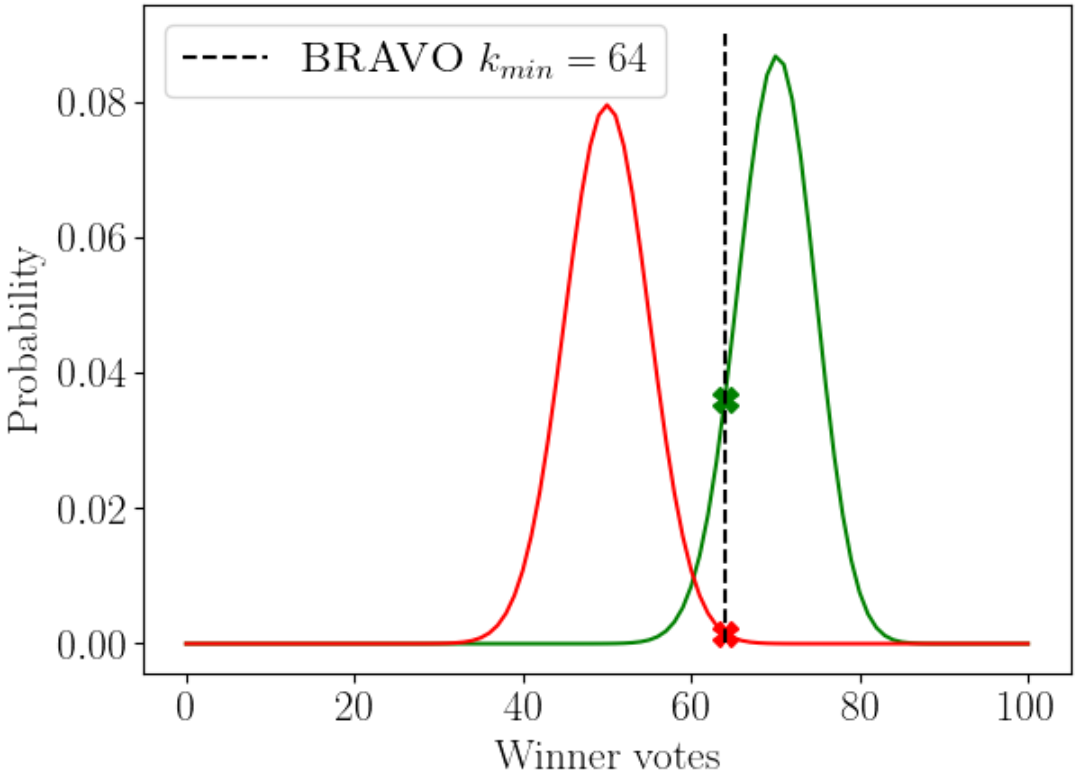

Figure 2.1: Probability distributions over the number of winner ballots K_1 in a first round of size n_1 for our toy example. The rightmost (green) curve corresponds to the alternative hypothesis (and thus is centered roughly at 0.7), whereas the leftmost (red) curve corresponds to the null hypothesis (and is centered at 0.5).

are drawn in the first round, and a new round of n_2 ballots are to be drawn. MINERVA proceeds by computing a new distribution: the probability that $K_2 = k_2$ and that the audit has not previously stopped, assuming the second round size is n_2. This distribution is the convolution of the a binomial distribution for the marginal, second round sample with the first round's distribution over those values of k for which the audit would not have stopped. Implicitly, this assumes that regardless of the value of K_1, the audit would still proceed to a second round of size n_2.

The MINERVA stopping rule is extended to future rounds in this way in order to give a nice property: the ratio of the stopping probability to risk in *every* round is bounded below by $\frac{1}{\alpha}$. If S_j is the stopping probability in round j (the probability that the audit stops in round j and no earlier, assuming the alternative hypothesis) and R_j is the risk in round j (the probability that the audit stops in round j and no earlier, assuming the null hypothesis), then the MINERVA stopping condition enforces the following constraint for all j:

$$\frac{S_j}{R_j} \geq \frac{1}{\alpha}$$

Equivalently, $R_j \leq \alpha S_j$. Note that $\sum_j R_j$ is the total risk of the audit and $\sum_j S_j$ is the total stopping probability of the audit. Taking the sum over all j, we get

$$\sum_j R_j \leq \alpha \sum_j S_j \implies \sum_j R_j \leq \alpha.$$

In other words, extending the tail ratio of MINERVA to future rounds in this way, so as to preserve this property of the audit, gives MINERVA its risk-limiting property. What we will see with PROVIDENCE is that this is not the only way to extend the tail ratio of MINERVA to future rounds while still preserving this property. Moreover, the MINERVA stopping condition achieves this nice property in a restrictive way; it assumes that next round sizes are constant, an undesirable property in practice.

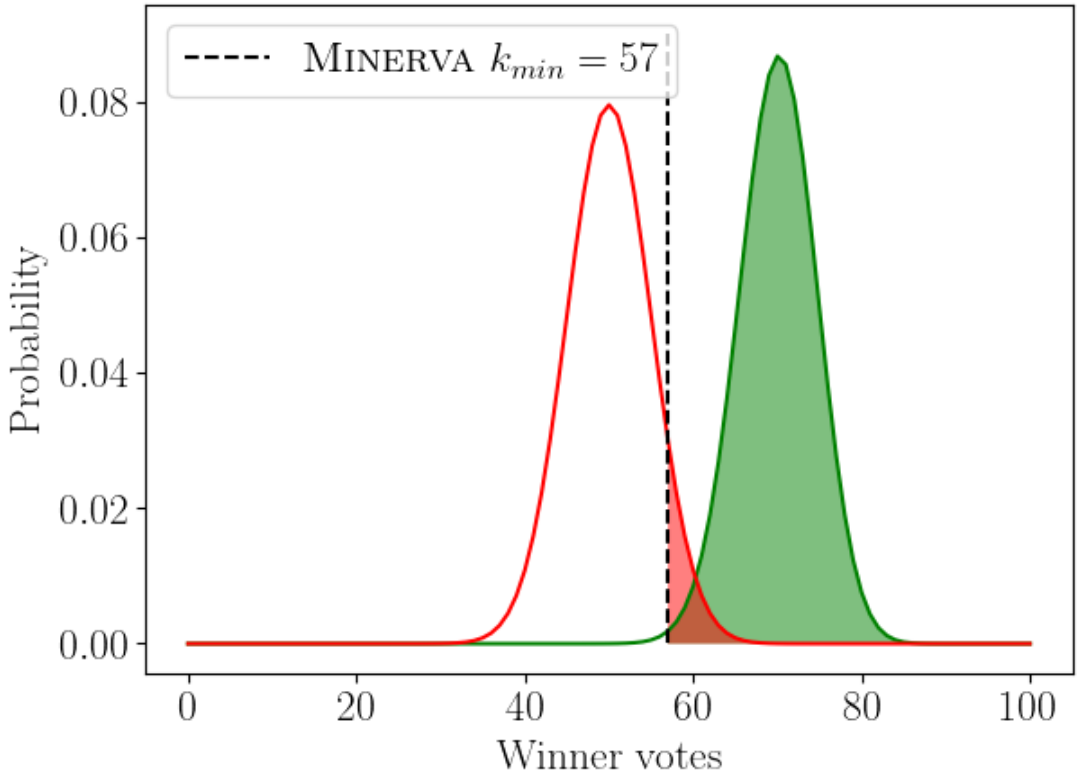

Figure 2.2: Probability distributions over the number of winner ballots K_1 in a first round of size n_1 for our toy example. The MINERVA value of k_{min} is shown, and the corresponding tails of the distributions shaded.

Chapter 3: PROVIDENCE

In this Chapter we describe the adversarial model and introduce the stopping condition of PROVIDENCE, proving some of its properties.

3.1 Adversarial model for RLAs

Detailed descriptions of best practices for post-election audits may be found in [9, 4]. For our purposes, we will assume that 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 are verified; and 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 in Chapter 1.1.1, we further assume that the ballot manifest is verified by the compliance audit; 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, because it can be checked by the public through knowledge of the seed and the drawn ballots. Thus the only variable is round size.

We define a *weakly round-choosing adversary* as one who can choose the first and subsequent round sizes before the audit begins and a *strongly round-choosing adversary* as one who can choose any round size at any time (before, of course, that particular round begins). In particular, a strongly round-choosing adversary may use information about the sample drawn thus far to decide the next round size, but a weakly round-choosing adversary

may not.

Definition 4 (Weakly Round-Choosing Adversary)**.** *A weakly round-choosing adversary may choose the first and consequent round sizes as a pre-determined function of audit parameters. That is, the j^{th} round size is a function*

$$n_j(\alpha, p_a, p_0, ballot\_manifest)$$

determined before the audit begins.

Definition 5 (Strongly Round-Choosing Adversary)**.** *A strongly round-choosing adversary may choose any round size as any function of audit parameters and all preceding samples. That is, the first round size is a function*

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

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

$$n_j(\alpha, p_a, p_0, ballot\_manifest, \boldsymbol{k}_{j-1}, \boldsymbol{n}_{j-1})$$

The functions n_j, $j \geq 1$ may be chosen at any time before the j^{th} round begins.

3.2 Intuition behind the properties of PROVIDENCE

Before defining PROVIDENCE, we give some intuition for how it is designed to avoid the problem of MINERVA.

Consider any ballot polling audit in round j, designed to stop for some value of $K_j = k_j$ smaller than the round size. In general, if k_j is large, the probability of this value of K_j given a correct outcome, $Pr[K_j = k_j \mid H_1]$, is larger than the corresponding risk—$Pr[K_j = k_j \mid H_0]$, the probability of this value of K_j given a tie. The risk is generally not zero, however. Thus,

corresponding to each such value of K_j, 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 may provide relationships among the various quantities. To obtain the risk of the entire audit, one sums the risks of each round.

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 which the audit stops, 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, expressions relating the risks are not available, and we are not able to obtain an expression bounding the sum of the risks across rounds. Thus we are not able to determine if MINERVA is an RLA in this case, or if it is vulnerable to the strongly round-choosing adversary [29, 28, 20].

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 a strongly round-choosing adversarial auditor determines round sizes after drawing the sample, and next round size computations may use knowledge of the current sample.

3.3 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, where samples are drawn with replacement, the R2* PROVIDENCE *stopping rule for the j^{th} round is:*

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

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

$$\begin{aligned} &\omega_j(k_j, k_{j-1}, p_a, p_0, n_j, n_{j-1}) \triangleq \\ &\sigma(k_{j-1}, p_a, p_0, n_{j-1}) \cdot \tau_1(k_j - k_{j-1}, p_a, p_0, n_j - n_{j-1}) \end{aligned} \tag{3.1}$$

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 (the distributions are binomial because the sampling is with replacement) 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. Smaller computational complexity makes audit planning and analysis using simulations as in Chapter 6 more feasible.

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

3.4 Proof of the risk-limiting property

We now prove that PROVIDENCE is risk-limiting against a strongly round-choosing adversary using lemmas from basic algebra which are given in Appendix A.

Theorem 1. $(\alpha, p_a, p_0, k_{j-1}, n_{j-1}, n_j)$-PROVIDENCE *is an* α*-RLA in the presence of a strongly round-choosing adversary.*

Proof. Let $\mathscr{A} = (\alpha, p_a, p_0, k_{j-1}, n_{j-1}, n_j)$-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 $\mathscr{A} =$ 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 7 and Definition 7, there is a value $k_{min,1} = k_{min,1,0,n_1}^{p_a,p_0,\alpha,0}$ 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 7, $\mathscr{A}$ = 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 8 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 7 and Definition 6, we see that there exists a $k_{min,j} = k_{min,j,n_{j-1},n_j}^{p_a,p_0,\alpha,k_{j-1}} \leq k_j$ 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[1] holds even if the values

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

of n_j depend on k_{j-1}, if n_j is a function $n_j(\alpha, p_0, p_1, ballot\_manifest, \boldsymbol{k}_{j-1}, \boldsymbol{n}_{j-1})$.

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)[2]. When both sides of the inequality are further summed over all rounds, we get:

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

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

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

□

3.5 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. For a conditional stopping probability of 0.9 in each round of a PROVIDENCE audit, we can compute next round sizes based on the current sample. MINERVA, however, would have a predetermined round schedule. We use the default MINERVA round schedule of audit software Arlo: $[x, 2.5x, 6.25x, ...]$; that is, the next marginal round size is 1.5 times the current one—equivalently, the next total round size is 2.5 times the current one. This approach is known to give, over a wide range of margins, a conditional probability of stopping roughly 0.9 in the second round if the first round size was determined for a probability of stopping 0.9.

Both the audits of our toy example therefore begin with a first round size of $17,272$ with

PROVIDENCE is resistant to a strongly round-choosing adversary.

[2]In fact, this is the relationship MINERVA enforces for its stopping condition, additionally requiring that n_j be fixed for all values of k_{j-1}.

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. The audits are identical in the first round, and both numbers of ballots may be computed as described in [29] and used by Arlo. 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 0.954. 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 first round 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.

3.6 Theoretical properties

Here we present two interesting theoretical properties of PROVIDENCE.

3.6.1 Efficiency

It is easy to see that, for any sample for which the EoR BRAVO stopping condition is met, the PROVIDENCE stopping condition is also met. This implies that PROVIDENCE will never draw more ballots than EoR BRAVO.

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. 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 [28, Theorem 6]. Note that we apply this result on τ_j for just $j = 1$.

□

3.6.2 Markov-like stopping condition

First we introduce useful notation for the observation made in this section.

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*

$$k_{min,j,n_{j-1},n_j}^{p_a,p_0,\alpha,k_{j-1}} = \min\left\{k : \omega_j(k, k_{j-1}, p_a, p_0, n_j, n_{j-1}) \geq \frac{1}{\alpha}\right\}. \tag{3.2}$$

As shown in Lemma 7, 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.

We can now describe the Markov-like property of PROVIDENCE. After $j-1$ rounds, having drawn n_{j-1} cumulative ballots with k_{j-1} cumulative winner ballots, the PROVIDENCE

stopping condition in round j with size n_j is equivalent to the PROVIDENCE stopping condition for a second round where the first round size is $n_1 = n_{j-1}$ with $k_1 = k_{j-1}$ winner ballots and a second round size $n_2 = n_j$. The particular sequence of round sizes and winner ballots does not affect the PROVIDENCE stopping condition; only the previous cumulative round size and number of winner ballots matters. MINERVA does not have this property, but BRAVO also does.

Lemma 2. *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. This is easily observed. □

3.7 (α, δ)-PROVIDENCE

As described in Chapter 2.1.1, MINERVA audits can be modified to additionally enforce a bound on the likelihood ratio σ, yielding the ATHENA audit. We note that PROVIDENCE too can be modified in this way. For shorthand, we refer to the resulting RLA as (α, δ)-PROVIDENCE.

Definition 8 ($(\alpha, \delta, p_a, p_0, k_{j-1}, n_{j-1}, n_j)$-PROVIDENCE)**.** *An $(\alpha, \delta, p_a, p_0, k_{j-1}, n_{j-1}, n_j)$-*PROVIDENCE *RLA outputs Correct if and only if*

1. $(\alpha, p_a, p_0, k_{j-1}, n_{j-1}, n_j)$-PROVIDENCE *outputs Correct, and*
2. $\sigma(k_j, p_a, p_0, n_j) \geq \dfrac{1}{\delta}$.

Chapter 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 [5], 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[15]. 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). We use risk limit $\alpha = 0.1$, as is common in RLAs, see for example [16] and [7]. All trials have a maximum of five 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, if the tally were correct, virtually no audits would progress beyond five rounds given the large conditional probability of stopping in each round.

4.1 Tied elections

In the simulations of PROVIDENCE audits of a tied election, the fraction of audits that stop, as shown in Figure 4.1, 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 reported provide insight into stopping probability and number of ballots drawn when the election is as reported. Figure 4.2 shows that the stopping probabilities over the first rounds are near and slightly above 0.9 as expected, since

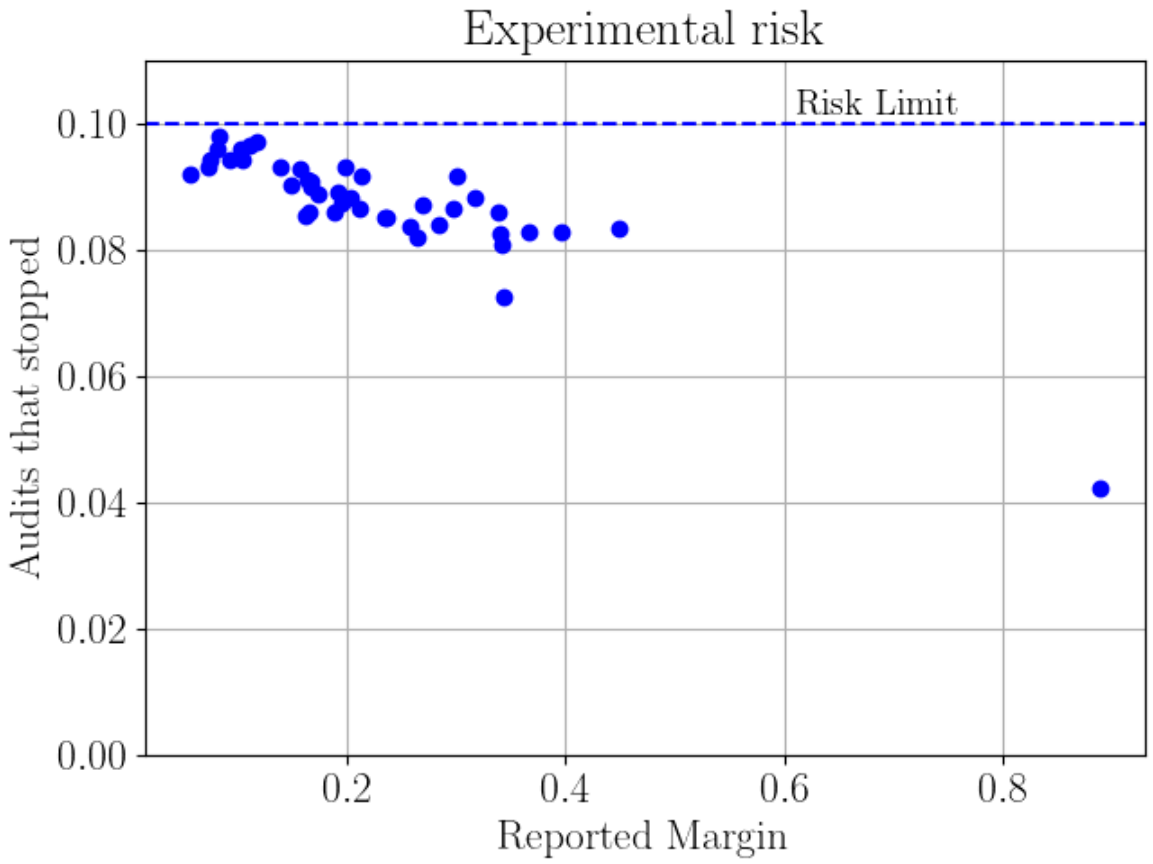

Figure 4.1: The fraction of simulated PROVIDENCE audits on tied elections that stopped in any round (we performed five rounds at a risk limit of 0.1) as a function of contest margin. This value is an estimate of the maximum risk of the PROVIDENCE audit. Observe that it is below the risk limit, as expected.

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.

4.2 Efficiency of PROVIDENCE

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

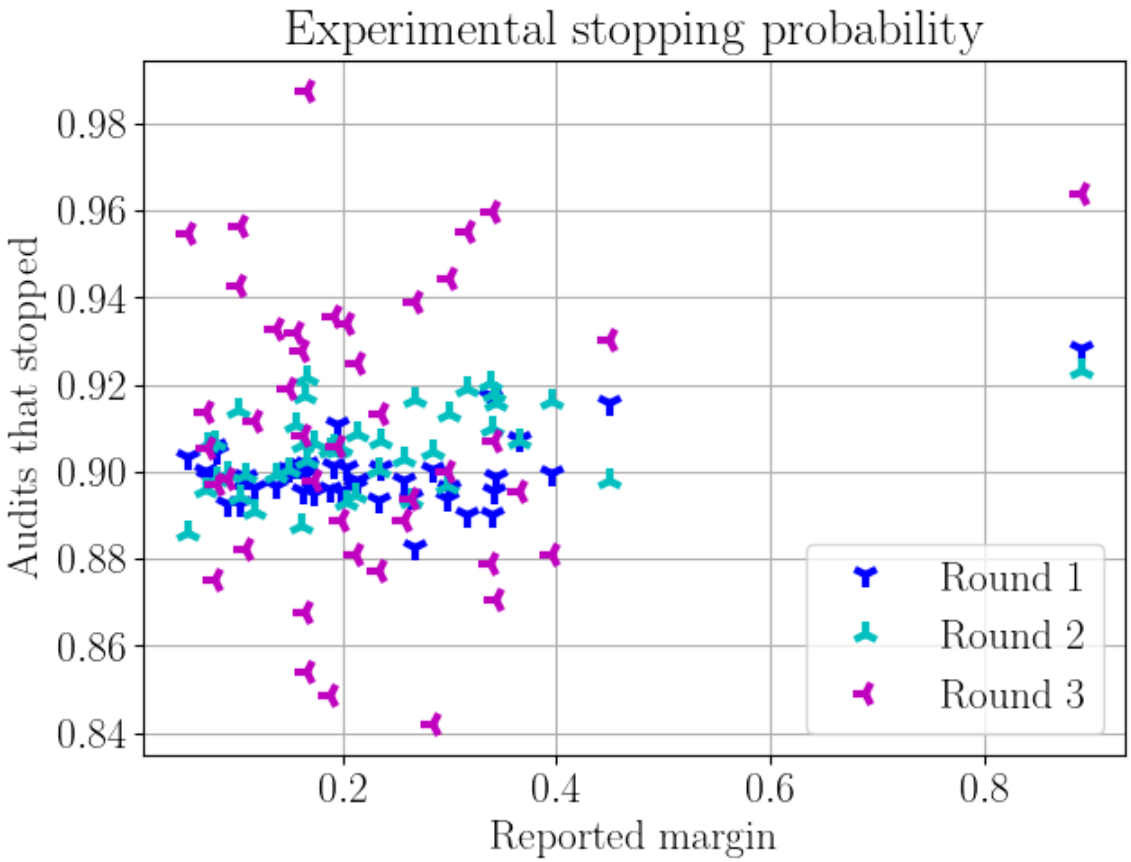

Figure 4.2: The fraction of simulated PROVIDENCE audits of the election as reported that stopped for each round as a function of margin. This value is an estimate of the stopping probability conditioned on the sample of the previous round. The average fraction for rounds 1, 2, and 3 is 0.8996, 0.9052, 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).

BRAVO implementations use a conditional stopping probability 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 4.3 shows the probability of stopping as a function of the number of ballots sampled, a plot similar to those presented in [5]. 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.

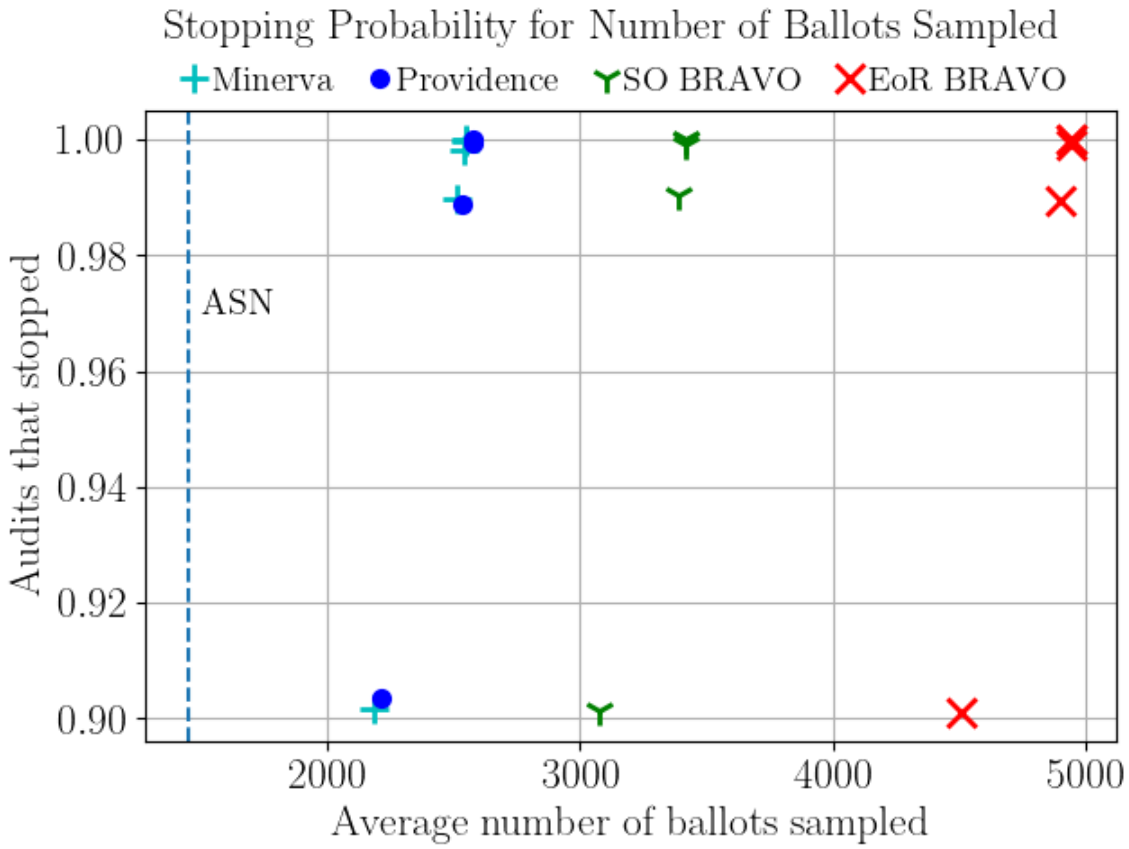

Figure 4.3: For the entire audit, consisting of all five rounds, the fraction of simulated audits that stopped as a function of the average number of ballots drawn for PROVIDENCE, MINERVA, EoR BRAVO, and SO BRAVO. The average sample number (ASN) for B2 BRAVO is included for context.

Chapter 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 measure was 0.0418. This is the smallest risk for which the sample would have passed a PROVIDENCE RLA for the announced election—the p-value of the statistical test and the inverse of the PROVIDENCE ratio ω of Definition 6. Table 5.1 shows risk measures for the drawn sample using PROVIDENCE, MINERVA and BRAVO (both EoR and SO), all similarly defined.

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

Table 5.1: Risk measures for the drawn first round of 140 ballots in the RI pilot audit. Risk measures in bold meet the risk-limit (10%) and thus correspond to audits that would stop.

Note that the risk measures shown in Table 5.1 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 5.1 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 of relevant votes drawn 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.

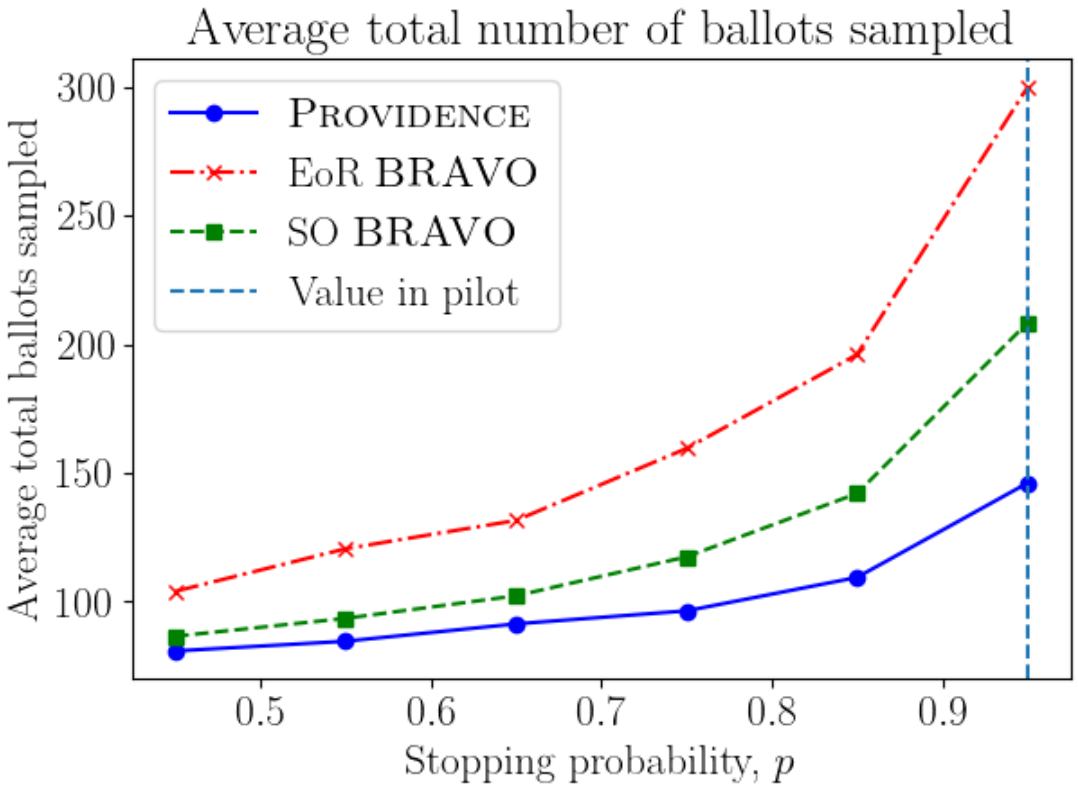

Figure 5.1: The total number of ballots sampled on average as a function of p, the conditional stopping probability used to select each round size. We use the same contest parameters and risk limit as the Rhode Island pilot.

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 Chapter 6 we consider a larger state-wide 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 Chapter 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 5.2 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 winner votes in the sample (orange line) surpasses the minimum number of winner ballots need to meet the BRAVO stopping condition (blue line) earlier in the sample[1]. EoR BRAVO, however, does not stop at sample size 140, the number of relevant ballots drawn during the pilot. 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 the rest of the sample is ignored. Second, the orange line is below the dotted green line, which represents half the sample size, at a sample size of 11; 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. Both of these observations are addressed more generally in Chapter 6.

We also compute the BRAVO test statistic, σ, for each cumulative number of winner votes throughout the selection order. Figure 5.3 shows that BRAVO ratio, σ from Equation 2.1. While the value of σ increases to above $1/0.1 = 10$ in the middle of the sample, it falls back below this line by the end.

[1]Such cases also provide insight into how PROVIDENCE is a sharper 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.

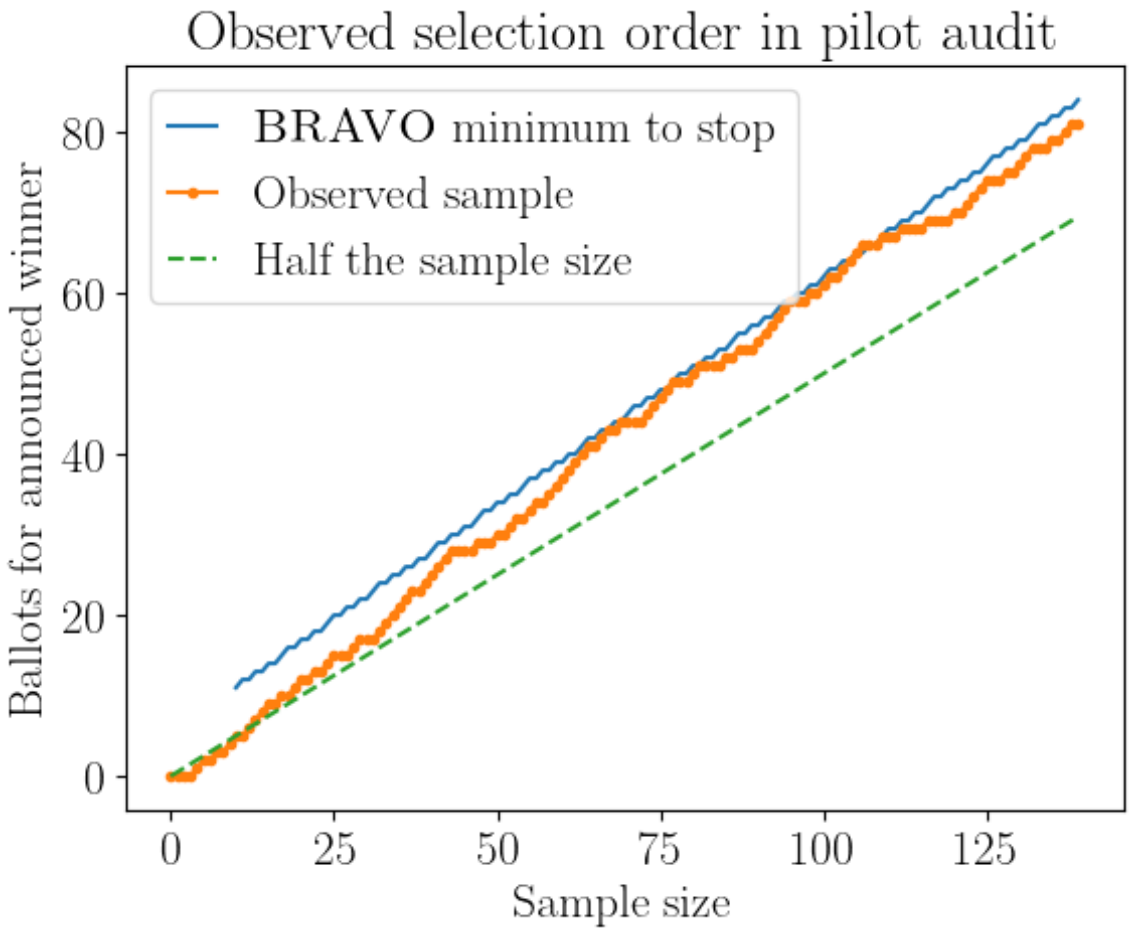

Figure 5.2: For each sample size from 1 to 140, the intermediate cumulative sum of ballots for the announced winner is shown. Observe that SO BRAVO would stop because the minimum number of ballots required to satisfy the BRAVO stopping condition (blue line) is achieved early on (when the orange line crosses the blue one) in the audit. However, because the BRAVO condition is no longer satisfied at 140 ballots, an EoR BRAVO audit would not have stopped. Further, at a sample size of 11, the orange line is not even above the dotted green line representing half the sample size and the winner has fewer than half the relevant votes in the sample.

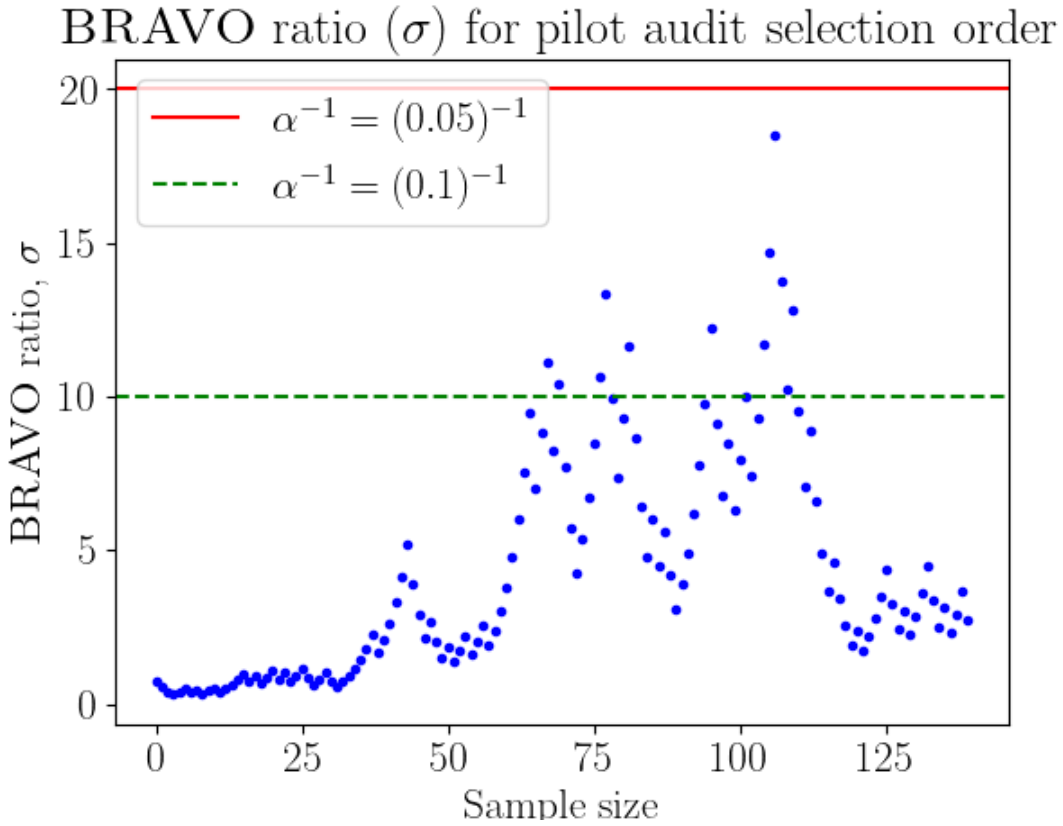

Figure 5.3: For each sample size from 1 to 140, the intermediate BRAVO ratio σ (Equation 2.1) is shown. Along the path, the ratio exceeds $1/0.1 = 10$ and so an SO BRAVO audit stops for this selection order at risk limit $\alpha = 0.1$, but $1/0.05 = 20$ is never exceeded and so an SO BRAVO audit with risk limit $\alpha = 0.05$ does not stop. Note that an EoR BRAVO audit does not stop for either risk limit because the condition is only evaluated at the end of the round.

Chapter 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 not straightforward. 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[1].

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[2]. Figure 6.1 shows the average total number of ballots sampled as a function of p.

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 6.1 shows 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

[1] 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.

[2] 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 very small[28]. It is straightforward to modify the model we discuss here to account for total distinct ballots.

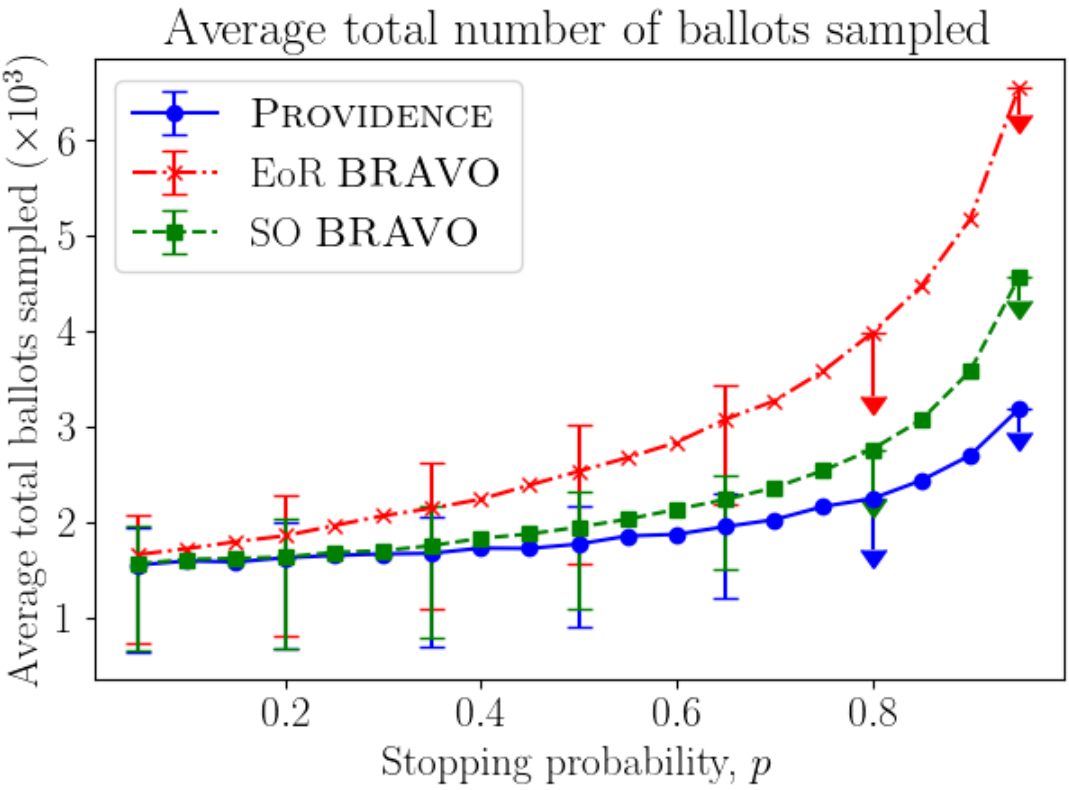

Figure 6.1: The average total number of ballots sampled, as a function of p, the conditional stopping probability used to select each round size, for ballot polling audits of the 2016 US Presidential election in the US State of Virginia. Error bars show the 0.25 and 0.75 quantiles. For sufficiently large p ($p \geq 0.75$), the 0.25 and 0.75 quantiles are both equal to the first round size, and this is shown by the downward arrows.

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 per-ballot workload w_b and a constant per-round workload c_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

$$W(E_b, E_r) = E_b w_b + E_r c_r + C \tag{6.1}$$

Note there is also some constant overhead of workload for the whole audit, namely C in Equation 6.1, 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 $c_r = x$ corresponds to a per round workload which is x times the per ballot workload. We use $c_r = 1000$ as a conservative example. That is, we set the overhead of a round equal to the workload of sampling 1000 ballots. Based on available data[7], the time retrieving and analyzing each individual ballot is on the order of 75 seconds which means that $c_r = 1000$ is equivalent to roughly 20 person-hours of workload. This corresponds to about 15 minutes being spent, on average, per round in each of the 133 counties of Virginia, a clearly conservative workload estimate. We do not consider $c_r < 1$ because it is not possible for the round overhead to be smaller than the workload corresponding to a single ballot.

As shown in Figure 6.2, 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 decrease 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 $c_r = 1000$.

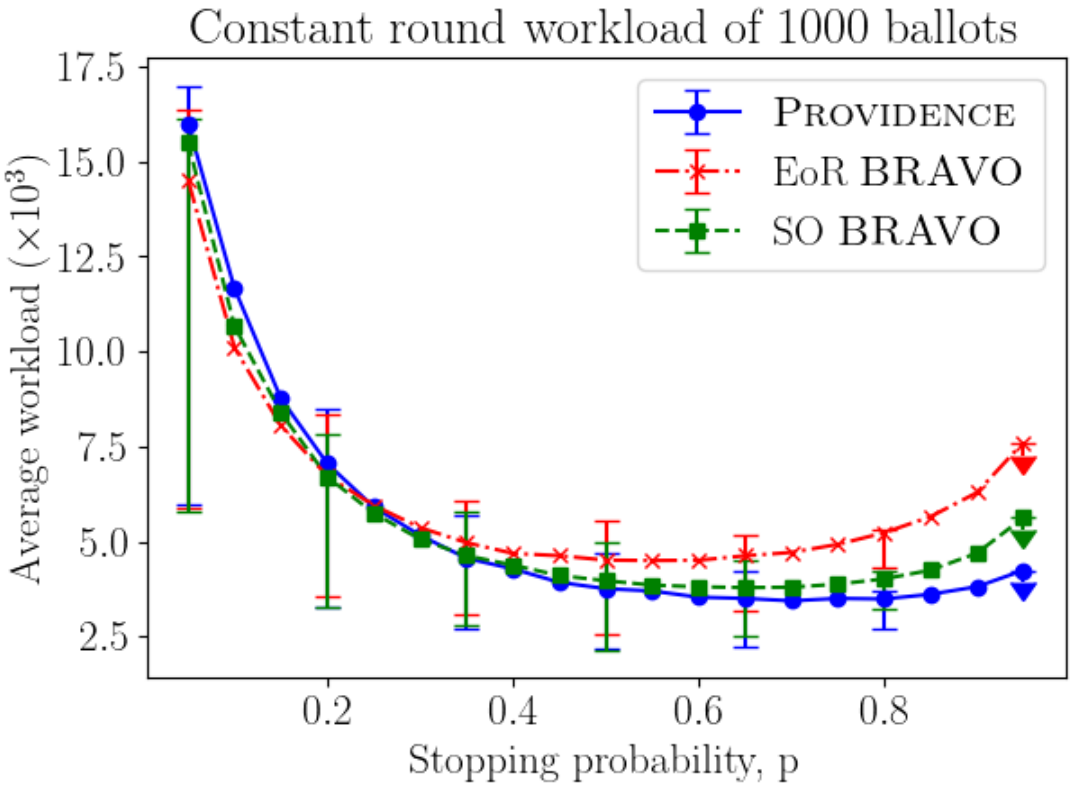

Figure 6.2: For workload parameters $w_b = 1$ and $c_r = 1000$, this plot shows the expected workload for various values of p. Expected workload is found using Equation 6.1 and the average number of ballots and rounds in our simulations as the expected number of ballots and rounds. The 0.25 and 0.75 quantiles are shown as in Figure 6.1.

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 c_r, we produce a dataset analagous to that of Figure 6.2 and then find the minimum average workload achieved for each of the audits and its corresponding stopping probability p.

Figure 6.3 shows the optimal achievable workload for a wide range of per round work-

loads. 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 c_r. We observe this behaviour in Figure 6.3 for extremely small and large workload values. For more reasonable values of the round workload c_r, SO BRAVO and EoR BRAVO achieve minimum workload roughly 1.1 and 1.3 times greater than that of PROVIDENCE.

Figure 6.4 shows the corresponding round schedule parameters p that achieve these minimal workloads. As expected, 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 c_r. Notice that PROVIDENCE is generally above and to the left of SO BRAVO, and SO BRAVO is generally above and to the left of EoR BRAVO. This relationship reflects the fact that for the same round workload, PROVIDENCE can get away with a larger stopping probability because it requires fewer ballots.

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

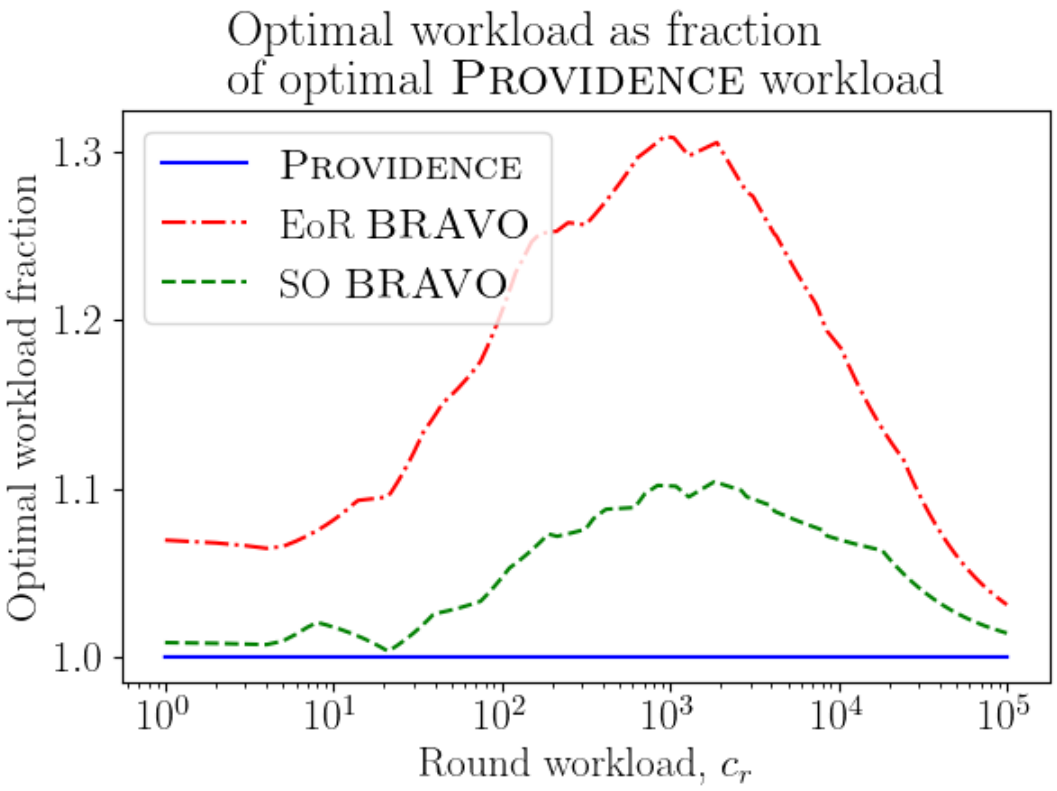

Figure 6.3: For varying round workload c_r, the optimal average workload achievable by each audit, as a fraction of the PROVIDENCE values.

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, c_p. In this model, the workload estimate incurs an additional workload of c_p every time a precinct is sampled from for the first time in a round. That is, let E_{pi} be the expected number of distinct precincts sampled from in round i, and let $E_p = \sum_i E_{pi}$. Then the new model is

$$W(E_b, E_r, E_p) = E_b w_b + E_r c_r + E_p c_p + C \tag{6.2}$$

We can again explore the minimum achievable workloads under this model, as shown in Figure 6.5.

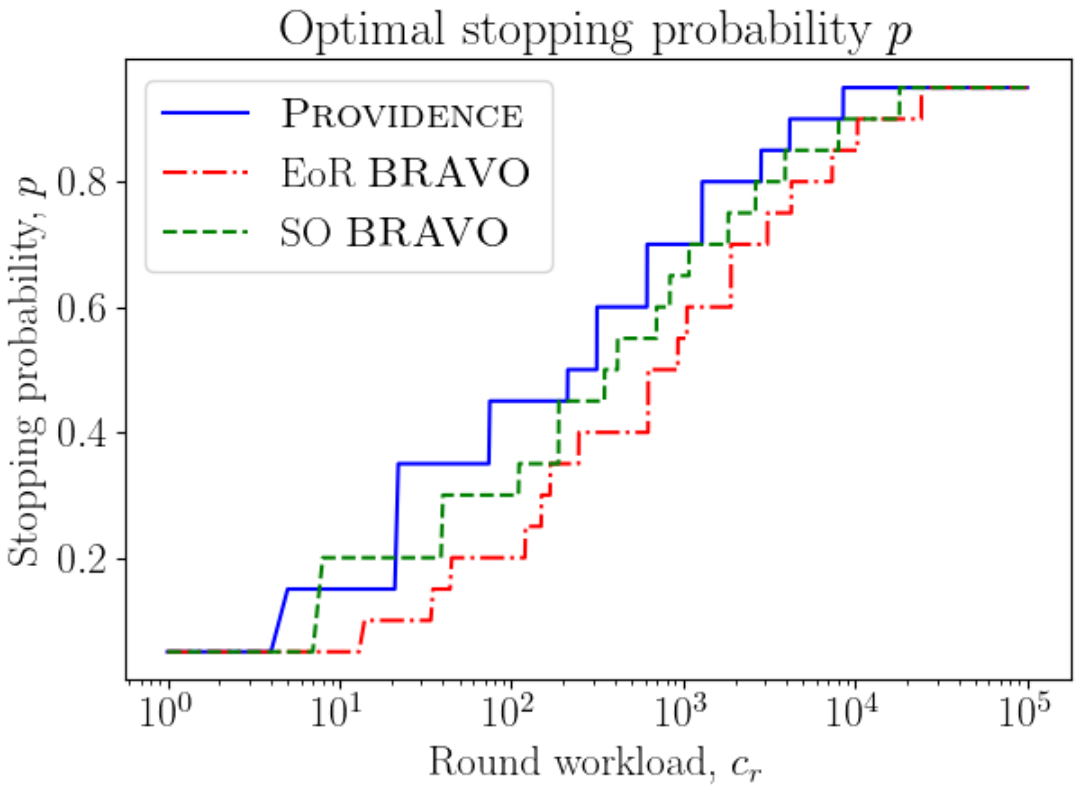

Figure 6.4: The optimal (workload-minimizing) stopping probability p for varying workload model parameters c_r. (Note that the steps in this function are a consequence of our subsampling the workload function. That is, the workload-minimizing value of p for each c_r is only allowed to take on values at increments of 0.05.)

6.2 Real time

Given tight certification deadlines, the total real time to conduct the RLA is also an important factor to consider when planning audits. Because each county can sample ballots for the same round concurrently, the total real time for a round depends only on the slowest county. In Virginia, Fairfax County typically has the most votes cast by a significant difference; in the contest we consider, Fairfax County had 551 thousand votes cast, more than double the 203 thousand of second-highest Virginia Beach City. Consequently, we model the expected total real time T of an audit using just the largest county, and we define analagous variables for the expected values in just the largest county. Note that some other

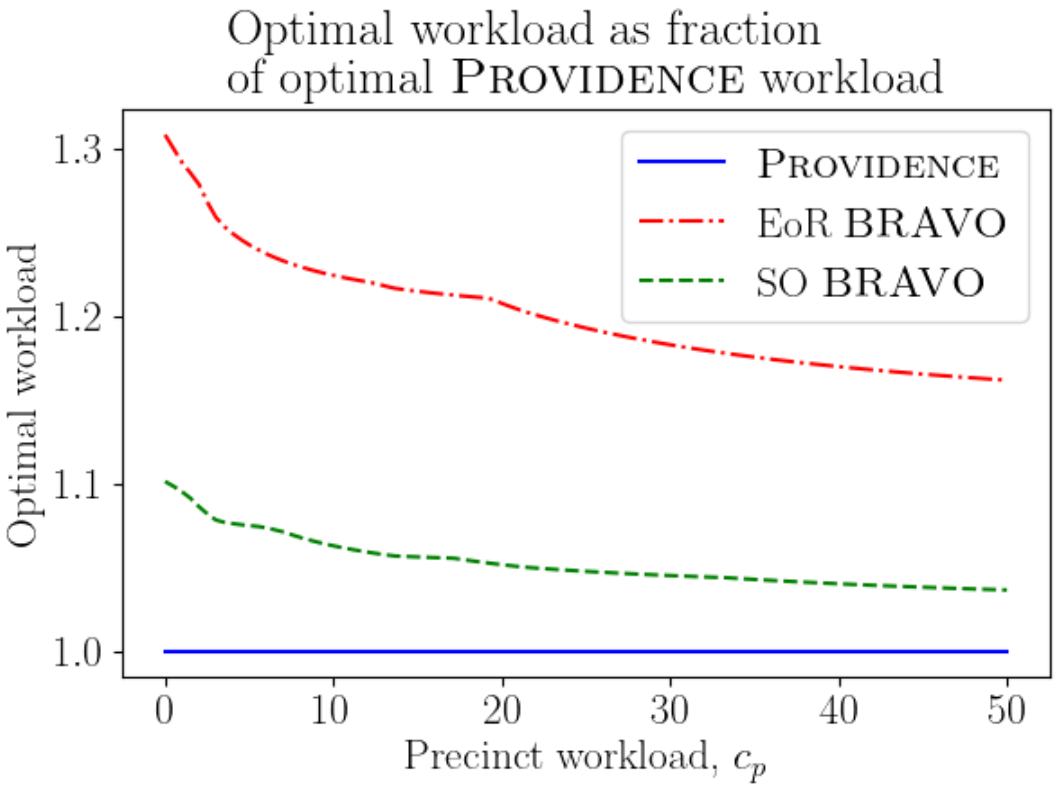

Figure 6.5: Optimal average workload using the workload Equation 6.2 for varying c_p, given as a fraction of the value for PROVIDENCE. Similar to Figure 6.3, we show a generous range of values for the workload variable, c_p in this case. If the time for a single ballot is 75 seconds, then $c_p = 50$ corresponds to over an hour of extra time to sample a ballot from a new container.

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

$$T(\bar{E}_b, \bar{E}_r, \bar{E}_p) = \bar{E}_b t_b + \bar{E}_r t_r + \bar{E}_p t_p + C \tag{6.3}$$

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 6.6 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 6.6 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 [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 of magnitude for this variable.

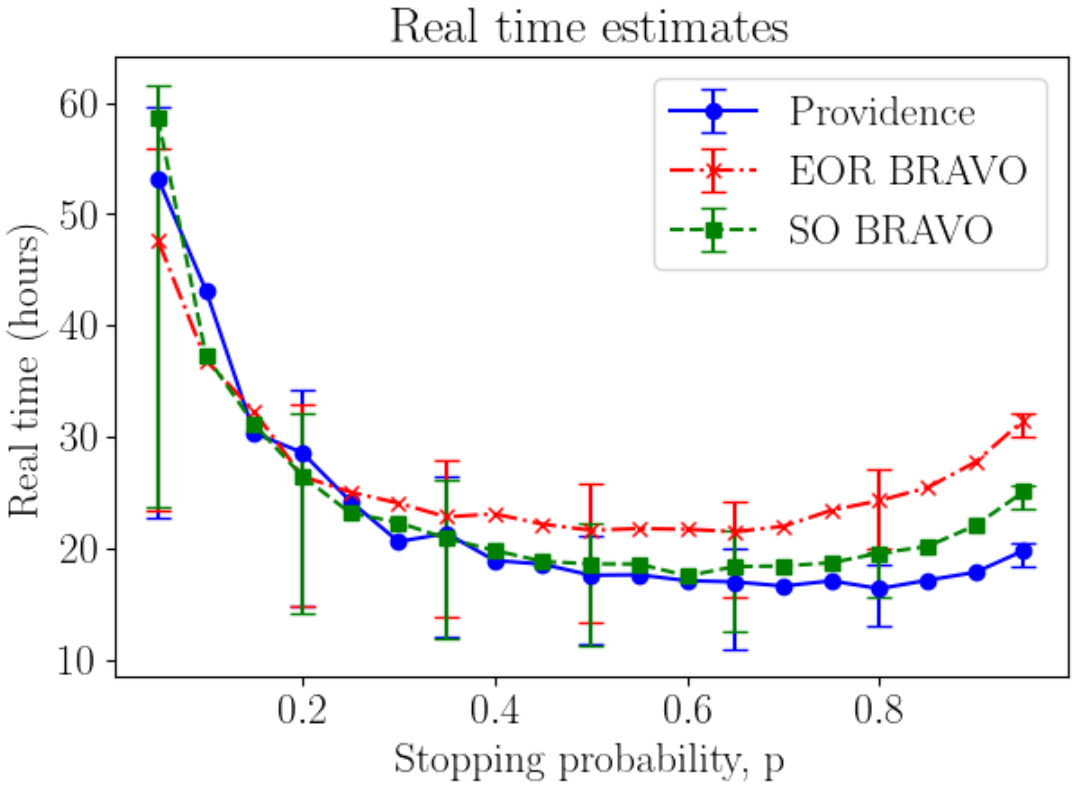

Figure 6.6: The real time as estimated by Equation 6.3 for varying p with expected values as estimated by our simulations. Error bars show the 0.25 and 0.75 quantiles. Unlike Figures 6.1 and 6.2, the quantiles still differ for large p because while the the number of ballots drawn in the first round in Virginia is constant, the number drawn in Fairfax County is variable.

Chapter 7: 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 5.2. 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 it serves to create expectations or suspicions that then need to be managed by election officials. Hence there is reason to structure round sizes so that they are unlikely to misrepresent the true outcome.

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 7.1 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 7.1, 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 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.

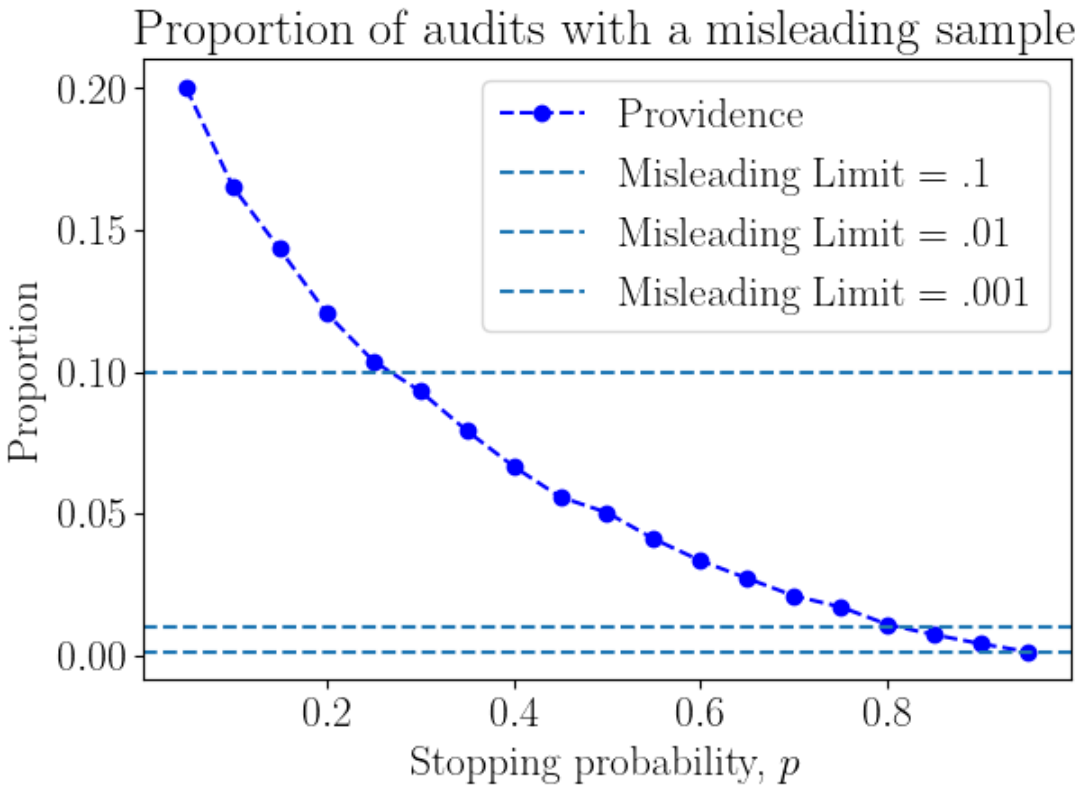

Figure 7.1: The proportion of simulated PROVIDENCE audits for the Virginia election parameters that had a *misleading sample* in any round.

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 7.1 also 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 7.1 shows for various margins the minimum first round size n that guarantees a

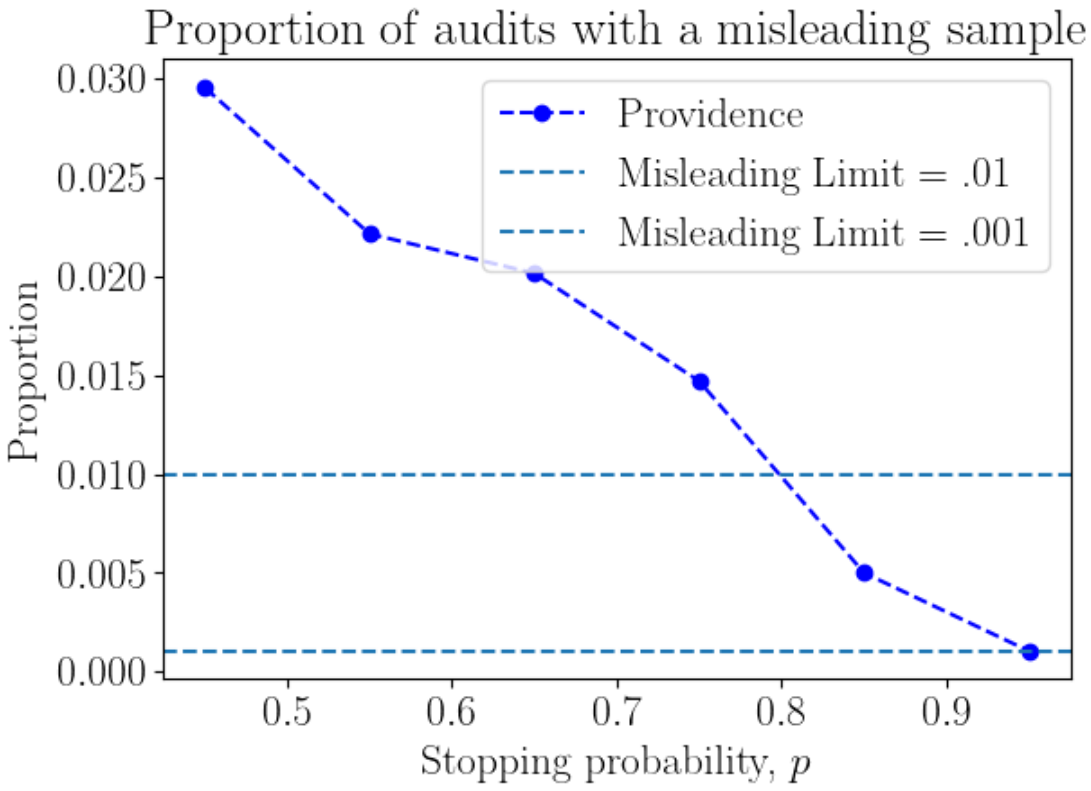

Figure 7.2: The proportion of simulated PROVIDENCE audits for the pilot audit parameters that had a *misleading sample* in any round.

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 7.1, 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.

7.1 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 [27]. 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.

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 7.1: 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.

In particular, for large round sizes n, there exist sequences which in total provide very strong evidence in support of the null hypothesis but which still stop, confirming the alternative hypothesis. While contrived (and unlikely), an example demonstrating such sequences is given in Figure 7.3. At best, this is an undesirable trait in the SO application of BRAVO in round-by-round audits which demonstrates its information inefficiency (stopping due to a sub-sequence and then ignoring later evidence). At worst, such a sequence could be drawn in an audit and leave election officials with a difficult job explaining the confirmation of reported results despite poor evidence.

Note that all possible selection orders with the same ultimate, cumulative tally of winner ballots occur with equal probability under both hypotheses. SO BRAVO accepts some such sequences and rejects other despite equivalent cumulative evidence.

Recall from Chapter 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 *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 **??** 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 7.4 only shows the proportion of audits that stopped with a *misleading sequence*; additional SO BRAVO audits also contained *misleading samples*.

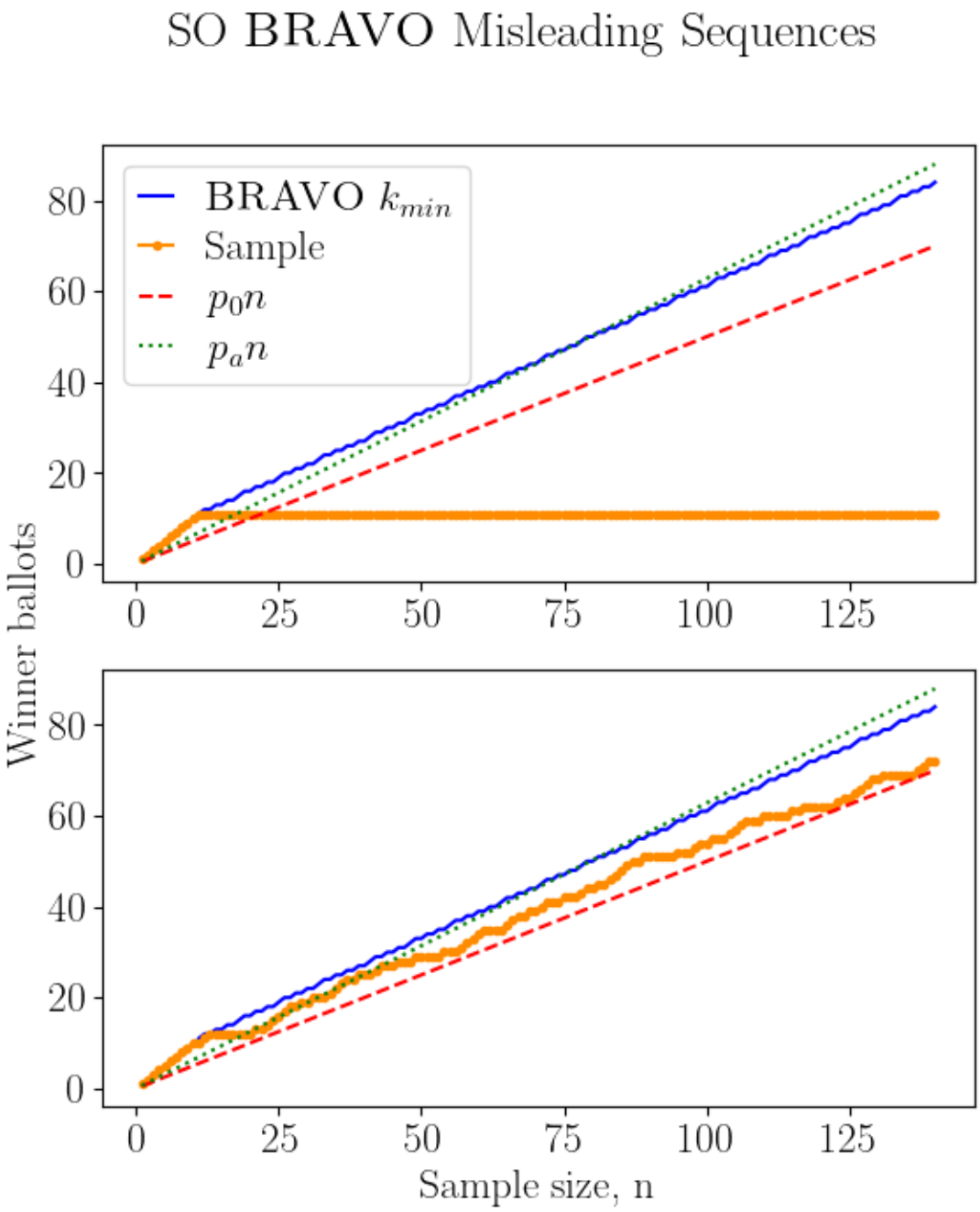

Figure 7.3: Contrived *misleading sequences* for which SO BRAVO audits would stop despite the cumulative sample at the end of the round providing very poor evidence for rejecting the null hypothesis. None of EoR BRAVO, MINERVA, and PROVIDENCE stop on these samples, yet SO BRAVO stops because of the early sub-sequence meeting the stopping condition; all later evidence in the sample is ignored. Note that sequences like the top example occur with negligible probability (it is shown instead to illustrate the information-inefficiency of SO BRAVO); the frequency of samples that meet the BRAVO stopping condition in an early sub-sequence but not at the end of a round is considered in Figure 7.4.

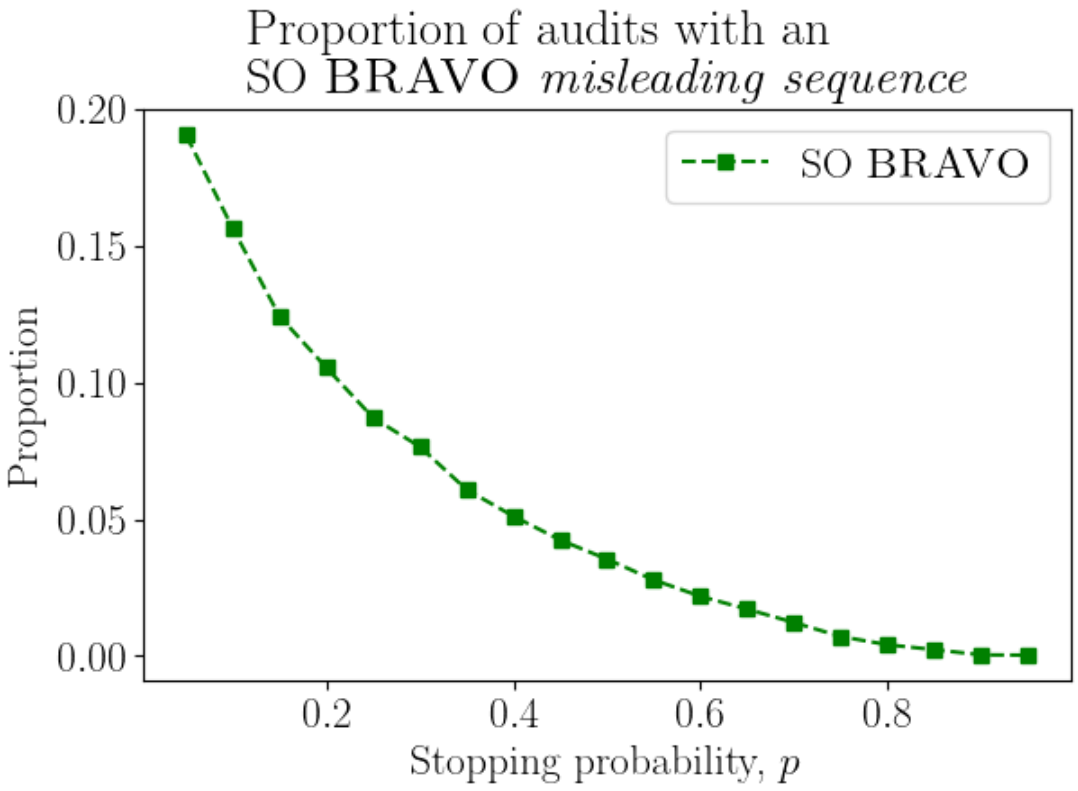

Figure 7.4: The proportion of simulated sequences that are misleading sequences in the SO BRAVO audit as a function of p.

Chapter 8: 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.

8.1 Availability

We provide an open source implementation of PROVIDENCE in the R2B2 software library for R2 and B2 audits [15]. The R2B2 implementation of PROVIDENCE has been incorporated as an option in Arlo, the most commonly used RLA software [25].

R2B2 also contains software to test stopping conditions and find round sizes for given probability of stopping and probability of a *misleading sample*. The code for simulations as well as workload and real time analysis is also provided.

8.2 Future work

8.2.1 Optimal ballot polling RLAs

As noted in Chapter 2, BRAVO is an instance of Wald's classic Sequential Probability Ratio Test (SPRT) [27]. Wald and Wolfowitz [26] showed that the SPRT is optimal in the sense that, among all sequential tests with power α, the SPRT minimizes the expected number of ballots sampled assuming the alternative hypothesis. In other words, for ballot-by-ballot audits, no other statistical test will generate more efficient ballot polling RLAs than the SPRT. Of course, we point out that real audits progress in rounds for a number of reasons.

Note that the SPRT deals with a special case of our problem, that with round schedule $[1, 1, 1, \ldots]$. For the more general round-by-round problem, one may wonder if there exists a test which, for any round schedule, minimizes the expected number of ballots sampled assuming the alternative hypothesis. It is unknown whether MINERVA or PROVIDENCE have this property. Either proving that one or both of MINERVA and PROVIDENCE have this optimality property or that some other test has it is a promising direction for future work.

Bibliography

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

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

[3] 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.

[4] 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: Why and how. `https://www.stat.berkeley.edu/~stark/Preprints/RLAwhitepaper12.pdf`, October 2012.

[5] 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.

[6] Oliver Broadrick, Poorvi L. Vora, and Filip Zagórski. Providence: a flexible round-by-round risk-limiting audit. In 32^{nd} *USENIX Security Symposium*, 2023.

[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] MIT Election Data and Science Lab. U.S. President 1976–2020, 2017.

[9] Lynn Garland, Mark Lindeman, Neal McBurnett, Jennifer Morrell, Marian Schneider, and Stephanie Singer. 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.

[10] 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.

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

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

[13] 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.

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

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

[16] Virginia State Board of Elections. 2022 risk-limiting audit report. `https://www.elections.virginia.gov/media/formswarehouse/risk-limiting-audit/2022-RLA-Final-Report.pdf`, March 2022.

[17] 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.

[18] 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.

[19] Ronald L. Rivest and John P. Wack. On the notion of “software independence” in voting systems. Prepared for the TGDC.

[20] PB Stark. ALPHA: Audit that learns from previously hand-audited ballots. *The Annals of Applied Statistics*, 2022.

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

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

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

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

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

[26] A. Wald and J. Wolfowitz. Optimum Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics*, 19(3):326 – 339, 1948.

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

[28] 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.

[29] 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.

Appendix A: Proofs

Lemma 3. *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 [29, Lemma 4]. □

Lemma 4. *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 [29, Lemma 2]. □

Lemma 5. *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 3-4. □

Lemma 6. *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 7. *For $\mathscr{A} = (\alpha, p_a, p_0, k_{j-1}, n_{j-1}, n_j)$-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*

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

Proof. From Definition 6,

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

Now to apply Lemma 6, 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 5. 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 5. Therefore by Lemma 6, we have the desired property. □

Lemma 8. *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

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

□