Simulations of Ballot Polling Risk-Limiting Audits

No Author Given

No Institute Given

Abstract. In this paper we present simulation results comparing the risk, stopping probability and number of ballots required over multiple rounds of ballot-polling RLAs Minerva, Selection-Ordered Bravo and End-of-Round Bravo. We also present details on the R2B2 open source software library and the related simulation software.
Bravo is the most commonly used ballot-polling risk limiting audit (RLA). Minerva was recently proposed, and requires fewer first-round ballots, on average, than both Selection-Ordered Bravo and End-of-Round Bravo when the first-round stopping probability is large.
An open question, however, is how Minerva compares to Selection-Ordered Bravo and End-of-Round Bravo over multiple rounds. In this paper, we present results from simulations of multiple round audits with first-round stopping probabilities of 0.9, a common choice among election officials. Because the size of rounds in Minerva needs to be predetermined (and independent of the audit draws), we use two pre-determined round sequences for Minerva: (a) all rounds are of the same size (which might be a practical choice for election officials if they had planned to have resources for the first round size) and (b) each round is 1.5 times the previous one (which is the preset value for Minerva as integrated into election audit software Arlo).
We show that the simulation results are consistent with predictions of the R2B2 open-source library for ballot polling audits. We also observe that both Bravo audits are unnecessarily conservative while Minerva audits stop with fewer ballots.

Keywords: Risk-limiting audit · Ballot polling audit

1 Introduction

1.1 Background

In the general ballot-polling risk limiting audit (RLA), a number of ballots are drawn and tallied in what is termed a *round* of ballots [**?**]. A statistical measure is then computed to determine whether there is sufficient evidence to declare the election outcome correct within the pre-determined risk limit. Because the decision is made after drawing a round of ballots, the audit is termed a *round-by-round (R2)* audit. The special case when round size is one—that is, stopping decisions are made after each ballot draw—is a *ballot-by-ballot (B2)* audit.

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

Minerva was designed for R2 audits and applies its stopping rule once for each round. Thus it does not require the tracking of ballots that SO Bravo does. It has been proven that Minerva is a risk-limiting audit and requires fewer ballots to be sampled than EoR Bravo when an audit is performed in rounds. Computations of first-round size for a 0.9 stopping probability when the election is as announced have been computed for a wide range of margins and shown to be smaller than those for both EoR and SO Bravo. First round simulations of Minerva [**?**] demonstrate that its first-round properties—regarding the probabilities of stopping when the underlying election is tied and when it is as announced—are as predicted for first round sizes with stopping probability 0.9. There are no results, either theoretical or based on simulations, regarding the expected number of ballots drawn in a Minerva audit. Further, there is no literature comparing simulations of Minerva and EoR or SO Bravo, or studying the sizes of multiple-round Minerva audits.

1.2 Our Results

1.3 Organization

2 Definitions

Now we present relevant definitions. We consider a two candidate plurality contest. To begin we define an audit.

Definition 1. *An audit $\mathcal{A}$ takes a sample X as input and gives one of the following decisions*

1. *Correct: the audit is complete*
2. *Uncertain: continue the audit*

All of the audits discussed in this paper are modeled as binary hypothesis tests. Under the alternate hypothesis, H_a, the announced outcome is correct. That is, the true underlying ballot distribution is given by the announced ballot tallies. Under the null hypothesis, H_0, the true outcome is a tie (or a the announced winner losing by one vote if there is an odd number of total ballots). Now we define key attributes of an audit that we will consider while analyzing our simulations. The risk of an audit is the probability that an audit stops when a tie has occurred.

Definition 2 (Risk). *The risk R of an audit $\mathcal{A}$ is*

$$R(\mathcal{A}) = \Pr[\mathcal{A}(X) = Correct \mid H_0]$$

This leads us to the following simple definition of an α-RLA.

Definition 3 (Risk Limiting Audit (α-RLA)). *An audit $\mathcal{A}$ is a Risk Limiting Audit with risk limit α iff*

$$R(\mathcal{A}) \leq \alpha.$$

It is useful to discuss the probability of an audit stopping in some round, if the outcome is correctly announced.

Definition 4 (Stopping Probability). *The stopping probability S of an audit $\mathcal{A}$ in round j is*

$$S_j(\mathcal{A}) = \Pr[\mathcal{A}(X_j) = Correct \mid H_a]$$

The notion of stopping probability can be useful for selecting round sizes.

3 Simulator

3.1 Simulations to Support Theoretical Audit Properties

The outcomes of RLAs depend on random chance; some random samples support the alternative hypothesis more than expected, resulting in quick low-risk conclusions, while other samples require subsequent rounds in order to confirm the announced results. We can simulate random samples for various underlying ballot distributions by computing pseudorandom samples. By applying an audit's stopping condition to thousands of such simulated samples, the average behavior of the simulated audits will tend towards the true behavior of the audit. In this way, we can examine whether theoretical claims about an audit are actually correct.

3.2 Software for Simulations

Our open source audit software library r2b2 [link] has implementations of several ballot polling risk-limiting audits as well as a simulator, all written in Python. For each of these audits, the software can evaluate the stopping condition for a given sample and can give estimates of the minimum round size to achieve a desired stopping probability. For a given audit and random seed, the simulator draws random samples using the pseudorandom number generator, [need to check]. Ballots can be sampled from any distribution of the users choosing. It is often useful to consider the distribution of ballots corresponding to a tie and the distribution of ballots corresponding with the announced results; these are the distributions represented by the null and alternative hypotheses. After drawing a sample, the simulator then evaluates the given audit's stopping condition for this simulated sample. If the audit stops, the simulation stops, and if the audit

continues, the simulation draws another round. The abstract simulator class does not prescribe any one method for choosing round sizes. We implement several classes to support various round size choices: round sizes from an estimate to achieve a desired probability of stopping, predetermined round sizes, and random round sizes.

4 Simulation Results

For this paper, we simulated audits for the 2020 Presidential election in all US states whose margin was at least 5%. Round sizes increase roughly proportional to the inverse square of the margin, so smaller margins are computationally much more expensive to simulate. For each of these states, we simulated $10,000 = 10^4$ audits with the announced underyling ballot distribution and an additional $10,000 = 10^4$ audits with a tie as the underlying ballot distribution. These are reasonable choices for inital simulation experiments because most audits frame the stopping decision as a binary hypothesis test where the null hypothesis assumes an underyling tie and the alternative hypothesis assumes the announced distribution. A standard first round size in ballot-polling audits has been one which achieves a 90% probability of stopping, and we chose round sizes to reflect this standard in our simulations. We ran our simulations for up to five rounds. With a 90% stopping probability, an audit only has a a probablity of not stopping by the end of the fifth round, assuming the outcome was correctly announced.

4.1 End-of-Round Bravo

For the EoR Bravo simulations, our software estimated and used for each round the minimum round size that would achieve a 90% probability of stopping. For the simulations with the announced outcome as the underlying ballot distribution, for each round j we computed the proportion of audits that stopped in round j among those that had not stopped before round j.

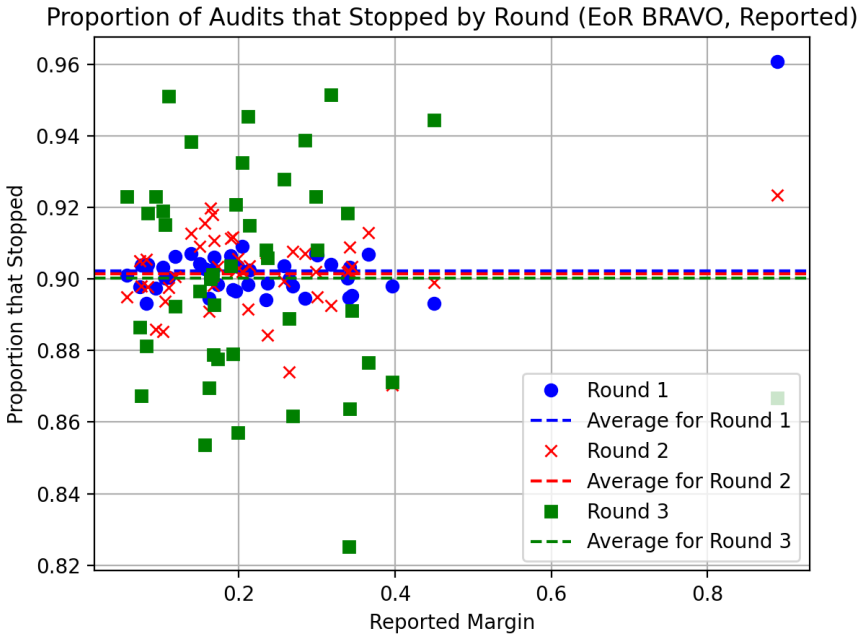

Fig. 1. For the first three rounds of the EoR Bravo simulations, this plot shows the proportion of audits that stopped in the jth round to all audits which had not yet stopped before the jth round.

In Figure 1, we display proportions for only the first three rounds since very few audits, $(.1)^{j-1} \cdot (10^4)$ on average, make it to the jth round. As a result, the proportions are based on an exponentially smaller dataset in each round. The proportions shown in Figure 1 give an estimate of the true probability of an EOR Bravo audit stopping in the jth round, given that it has not already stopped in a previous round. In 1, we see that, especially in earlier rounds for which the values are more representative of true audit behavior, our predictions are accurate. In particular, the average across all margins is just above $.9 = 90\%$ for all three rounds.

The risk of this audit, across all 5 rounds, is an important metric since it determines whether an audit is risk-limiting. Therefore, we now consider the proportion of audits that stopped with an underlying tie. This proportion, for a risk-limiting audit, should approach a value less than the risk limit, 0.1, as more audits are performed.

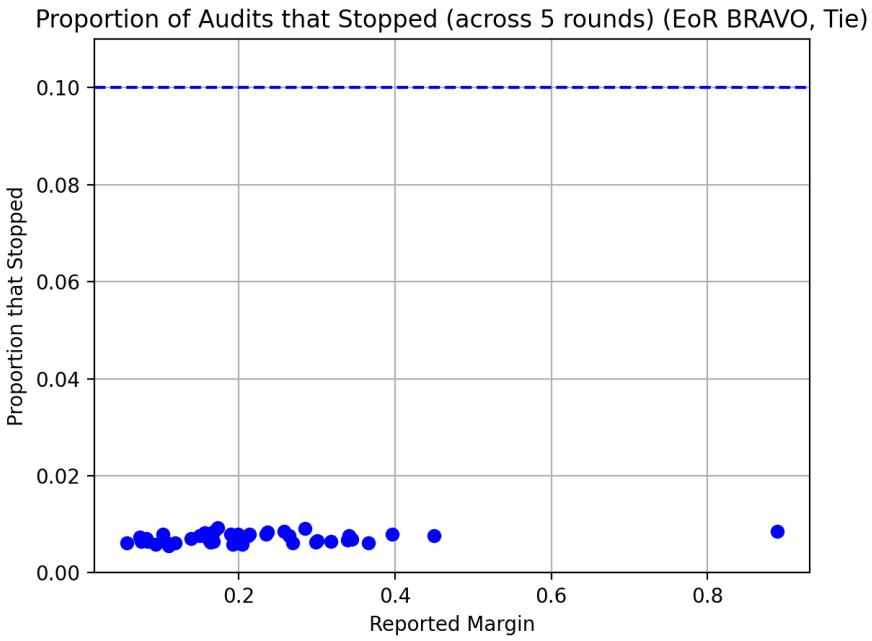

Fig. 2. For each state margin, this plot shows the proportion of EOR Bravo audits with an underlying tie that stopped.

Figure 2 makes it clear that the risk of EOR Bravo is roughly an order of magnitude less than the risk limit. That is, on average, roughly 10 times more audits could have stopped with the audit still meeting the risk limit. These simulations support the claim made in [Minerva paper] that EOR Bravo is unnecessarily conservative and thus requires more ballots on average.

4.2 Minerva Simulations

For Minerva, it has not been shown that round sizes can be chosen during the audit. That is, an adversary with knowledge of the history of the audit may be able to choose round sizes which cause the risk of the audit to exceed the risk limit. For this reason, we have to choose the round sizes of a Minerva audit a priori. For this paper, we consider two choices of round sizes. For both, we estimate and then use the minimum first round size which achieves a 90% probability of stopping. Then, for subsequent rounds, we either (i) draw the same number of ballots in each round or (ii) multiply the previous round size by a factor of 1.5 and sample this many new ballots. We consider the case of drawing samples of the same size because it may reflect a practical way to continue an audit; if election officials have selected some first round size within reasonable

logistical bounds, drawing the same number of ballots in subsequent rounds may be practical. We also consider round sizes with samples increasing by a multiple of 1.5 because this multiple gives a very rough approximation of round sizes with a 90% probability of stopping.

Round Sizes with Multiple of 1.0 As with the preceding simulations, we ran $10,000 = 10^4$ trials per state for both the underlying tie and underlying reported outcome.

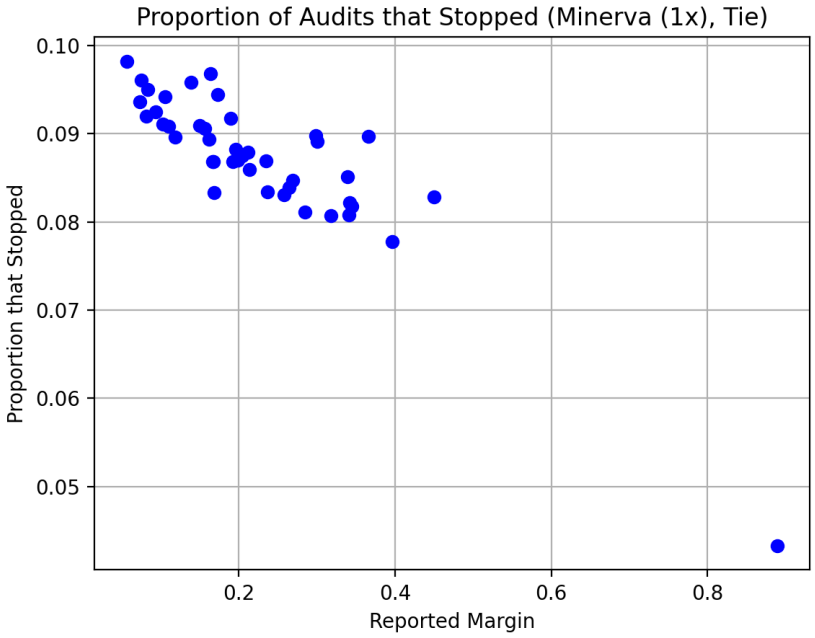

Fig. 3. This plot shows, for each state's margin, the proportion of audits that stopped of all 10^4 audits with an underlying tie in the simulations with a round size multiple of 1.0.

Notice in Figure 3 that all simulations for a tie had a proportion of audits that stopped less than .1, the risk limit, supporting the claim that Minerva is a risk limiting audit. Unlike EOR Bravo, the experimental risks here are much closer to the risk limit, showing that Minerva stops on average with a less conservative risk; Minerva is sharper.

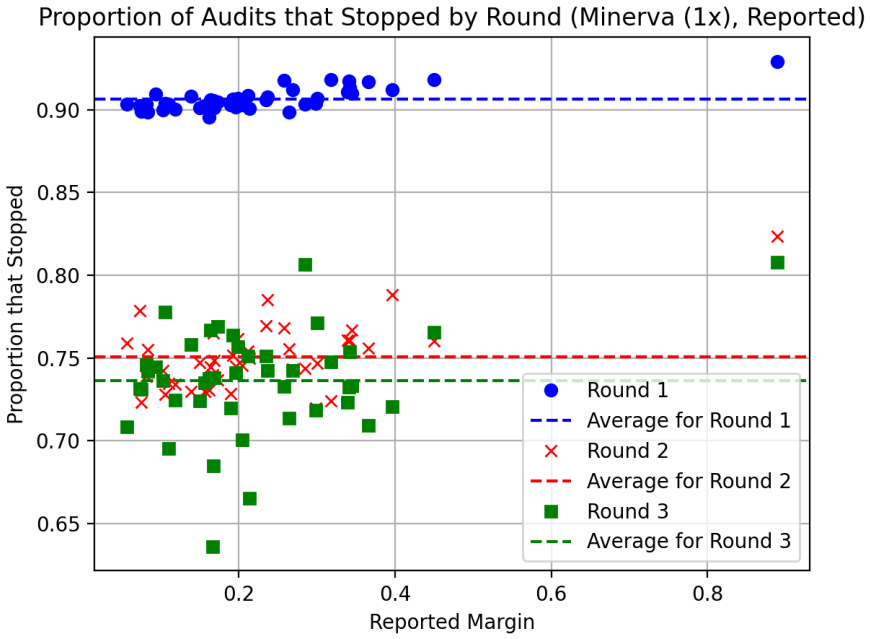

Fig. 4. This plot shows, for each state's margin, the proportion of audits that stopped of all 10^4 audits with the announced results as the underlying distribution in the Minerva simulations with a round size multiple of 1.0.

Figure 4 shows that the round size estimate for the first round achieved the desired stopping probability of 90%. For subsequent rounds, the multiplier of 1.5 did not consistently achieve 90% stopping probability, but it was within roughly 10%.

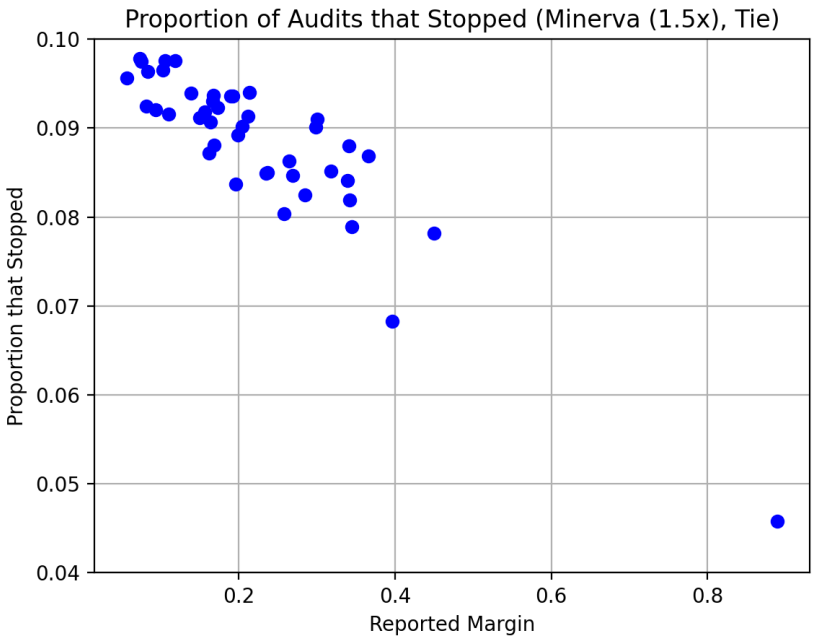

Fig. 5. This plot shows, for each state's margin, the proportion of audits that stopped of all 10^4 audits with an underlying tie in the simulations with a round size multiple of 1.5.

Round Sizes with Multiple of 1.5 Figure 5 shows, for each state's margin, the proportion of audits that stopped of all 10^4 audits with an underlying tie in the simulations with a round size multiple of 1.5.

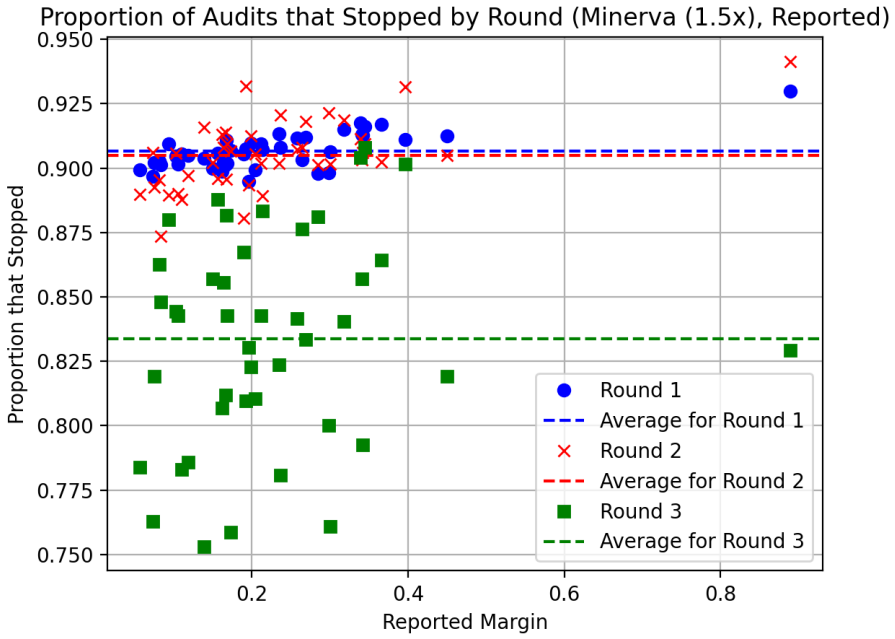

Fig. 6. This plot shows, for each state's margin, the proportion of audits that stopped in each round with the announced results as the underlying distribution in the MINERVA simulations with a round size multiple of 1.5.

Figure 6 shows that the round size estimate for the first round achieved the desired stopping probability of 90%. For subsequent rounds, the multiplier of 1.5 did not consistently achieve 90% stopping probability, but it was within roughly 10%.

References

1. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
4. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
5. LNCS Homepage, http://www.springer.com/lncs. Last accessed 4 Oct 2017