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 risk limiting audits (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 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 90%, 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 more conservative than needed, while Minerva audits stop with fewer ballots.

Keywords: Risk-limiting audit · Ballot polling audit

1 Introduction

The literature contains numerous descriptions of vulnerabilities in deployed voting systems, and it is not possible to be certain that any system, however well-designed, will perform as expected in all instances. For this reason, *evidence-based elections* aim to produce trustworthy and compelling evidence of the correctness of election outcomes, enabling the detection of problems with high probability. One way to implement an evidence-based election is to use a well-curated voter-verified paper trail, compliance audits, and a rigorous tabulation audit of the election outcome, known as a risk-limiting audit (RLA). This paper provides

insight into the main approach to ballot polling RLAs, the Bravo audit [8], and the newer Minerva [17] ballot polling RLA, through the presentation of simulation results. While some properties of the two audits may be theoretically derived, for other properties theoretical results are not available. This paper examines the number of ballots drawn over multiple rounds of both audits, and also examines the related probabilities of stopping if the election is as announced, and the maximum risk of the audit.

1.1 Background

This paper focuses on ballot-polling RLAs, which require a large number of ballots but do not rely on any special features of the election technology. 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 [17]. A statistical measure is then computed to determine whether there is sufficient evidence to declare the election outcome correct within the pre-determined risk limit. 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. Simulations have also been presented to understand better its first round properties; these show that Minerva also draws fewer ballots than SO Bravo for large first rounds. There are no results, however, either theoretical or based on simulations, regarding the expected number of ballots drawn over multiple rounds in a Minerva audit. Further, there is no literature comparing the number of ballots drawn by Minerva and EoR or SO Bravo over multiple rounds.

Both Bravo and Minerva have been integrated into election audit software *Arlo* [14], and, as such, available for use in real election audits. Both have been used in real election audits. For this reason, it is very important to understand their properties over multiple rounds.

1.2 Our Results

1.3 Organization

2 Related work

The Bravo audit [8] is a well-known ballot polling audit which has been used in numerous pilot and real audits. When used to audit a two-candidate election, it is an instance of Wald's sequential probability ratio test (SPRT) [15], and inherits the SPRT property of being the most efficient test (requiring the smallest expected number of ballots) if the election is as announced. The model for Bravo and the SPRT is, however, that of a sequential audit: a sample of size one is drawn, and a decision of whether to stop the audit or not is taken. Real election audits invest in drawing a large number of ballots before making the decision. It is possible to apply Bravo to the sequence of ballots if the sequential order is retained. This is not, however, the most efficient possible use of the drawn sample because information in consequent ballots is ignored when applying Bravo to a set of ballots drawn earlier.

We do know a great deal about the properties of Bravo. The risk limiting property of Bravo follows from the similar property of the SPRT. Stopping probabilities for Bravo may be computed as described by Zagórski *et al.*[17,16]. The results of the computations match simulation results reported by Lindeman *et al.* [8, Table 1].

The Minerva audit [17,16] was developed for large first round sizes which enable election officials to be done in one round with large probability. It uses information from the entire sample, and has been proven to be risk limiting when the round schedule for the audit is determined before the audit begins. That is, information about the actual ballots drawn in the first round cannot be used to determine future round sizes. First-round sizes for a 0.9 stopping probability when the election is as announced have been computed for a wide range of margins and shown to be smaller than those for both EoR and SO Bravo. First round simulations of Minerva [16] demonstrate that its first-round properties—regarding the probabilities of stopping when the underlying election is tied and when it is as announced—are as predicted for first round sizes with stopping probability 0.9.

Ballot polling audit simulations have been used to familiarize election officials and the public with the approach [12]. McLaughlin and Stark [10,9] compare the workload for the Canvass Audits by Sampling and Testing (CAST) and Kaplan-Markov (KM) audits using simulations. Blom *et al.* demonstrate the efficiency of their ballot polling approach to audit instant runoff voting (IRV) using simulations [6]. Huang *et al.* present a framework generalizing a number of ballot polling audits and compare their performance (round sizes and stopping probabilities) using simulations [7]. This work was prior to the development of Minerva, and focuses on the comparison between Bayesian audits [11] and Bravo; essentially studying the impact of the prior of the Bayesian RLA.

3 Software

In this section we describe the software implementing ballot polling audits, termed the R2B2 library, and the simulator software used for this research. All the software is released as open source under the MIT License.

3.1 R2B2 Library

The R2B2 Python library [5] provides a framework for the exploration of round-by-round and ballot-by-ballot RLAs. The goal in designing R2B2 is two fold:

1. Provide an elegant Python library which can be easily imported and used in any other code base.
2. Provide an interactive set of tools which can be utilized 'out-of-the-box' for experimenting with and learning about risk-limiting audits.

Design The high-level design of R2B2 is an object-oriented view of election audits. The three main object classes, `Election`, `Contest`, and `Audit`, serve to group data into logically independent structures.

The `Election` contains the information that comprises an entire election, most importantly, the total number of ballots cast in the election and the list of `Contest`s from the election. At the moment `Election` does not offer functionality beyond grouping `Contest`s.

The `Contest` contains the information related to a single contest such as the ballots cast in that contest, the candidates, the type of contest, and the reported tally. Providing a structure to hold this information independent of any particular audit is especially useful for exploratory work.

The `Audit` contains information related to the audit parameters for a single contest, such as the risk limit, sampling method, and `Contest` to audit. It is important to note the `Audit` is an Abstract Base Class upon which specific RLAs are built. It only contains the parameters and attributes common to the RLAs of this paper and provides a set of methods that can be called by any audit implementation. The functionality of `Audit` can be divided into two basic groups: *interactive* and *bulk*.

The interactive implementation allows users to execute an audit step-by-step as it might progress during a live election audit through the following:

- The `run()` method begins an interactive audit executing where users are prompted for round sizes and the counts of winner ballots found in the sample and in return are given information about the current state of the audit and whether the stopping condition(s) have been met.
- Two distributions representing the null and alternative hypotheses are maintained and allow for computation of the audits per-round risk and stopping probability schedules.

– Before each round, the audit will recommend possible next round sizes given different criteria, such as a set of desired stopping probabilities.

The bulk implementations allows users to generate a larger set of data from an audit such as:

– A set of stopping conditions given a set of round sizes.
– A set of risk levels given a set of round size and winner ballots pairs.
– A list of all stopping conditions from the minimum to the maximum round size.

Usage R2B2 makes understanding and exploring election audits simple for the user with no Python knowledge while simultaneously providing a comprehensive set of tools for the experienced Python developer.

Using R2B2 is as simple as using any other Python library: simply import the library and all of the functionality is at your finger tips. Not only does this allow users to write their own Python scripts for exploring RLAs, it also allows R2B2 to be plugged in to any other Python library. See the following Jupyter Notebooks for information on the usage of R2B2: Basic Usage [3], Generating Graphs [4].

R2B2 also provides a significant amount of functionality 'out-of-the-box' for educational or exploratory use. For those who wish learn about RLAs without having to write any code themselves, R2B2 provides a command line tool for both interactive auditing and generating audit results and statistics for larger data sets.

3.2 Simulation Software

As described above R2B2 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 compute the stopping condition for a given sample and estimates of the next round size to achieve a desired stopping probability. For a given audit and random seed, the simulator draws random samples, with replacement, using a pseudorandom number generator, [need to check]. given the number of votes for each candidate, and the number of invalid votes, in the underlying election (these need not be chosen to be those announced).

When the number of candidates is more than two, the audit is carried out pairwise for each candidate pair, and votes for all other candidates are considered invalid votes.

After drawing a simulated sample of ballots, the simulator evaluates the given audit's stopping condition for this 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 pseudorandomly-generated round sizes.

3.3 Testing

The R2B2 software is used to compute stopping conditions and next round estimates. It is intended for use by us and other researchers, and designed for this purpose. We have also independently implemented all the functionality in matlab [2] (the two codebases are written by different individuals) and have extensively checked the results of the two codebases. Additionally, for use in regular election audits by election officials, we have written an add-on [1] to the *Arlo* risk-limiting audit software, the results of which have also been extensively checked against the other two codebases.

4 Experiments

In this section, we motivate and describe the experiments. We consider a two candidate plurality contest, and assume that ballots are sampled with replacement, as is common in the literature.

We first present relevant definitions.

Definition 1. *An audit $\mathcal{A}$ takes a sample of ballots 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 the announced winner lost by one vote, and the number of ballots is large enough that the probability of drawing a ballot for the winner is that of drawing one for the winner).

The maximum risk of an audit is the probability that an audit stops, given that the underlying election is a tie. (Vora show that this is the maximum risk [13].)

Definition 2 (Risk). *The maximum 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 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 the j^{th} round, given that the underlying election is as announced.

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

$$S_j(\mathcal{A}) = \Pr[\mathcal{A}(X) = \textit{Correct in round } j \ \wedge \mathcal{A}(X) \neq \textit{Correct previously} \mid H_a]$$

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

We performed simulations to study S_j and obtain estimates of the expected number of ballots drawn for the various audits. We know that Bravo uses the smallest expected number of ballots when drawn ballot-by-ballot. However, this may not be the case when Bravo is used as an R2 audit for a large first round size. We also studied the risk values. While all the audits we studied have been proven to be RLAs, it is useful to observe how close to the risk limit each audit gets over a small, finite number of rounds as would be the case in a real election.

5 Simulation Results

For this paper, we simulated audits for the 2020 Presidential election in all US states where the pairwise margin for the two main candidates 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 assuming the underlying election was as announced, and an additional $10,000 = 10^4$ audits assuming the underlying election was a tie.

We ran simulations for a 90% probability of stopping in the first round, enabling election officials to be done in the first round with very high probability if the election is as announced. We ran our simulations for up to five rounds. We assume that the audit results in a hand count if it does not stop in any of 5 rounds.

5.1 End-of-Round Bravo

For the EoR Bravo simulations, our software estimated and used for each round the round size that would achieve a 90% probability of stopping if the true election outcome were as announced. In Figure 1, we display the proportion of EoR Bravo audits that stopped in the j^{th} round to all audits which had not yet stopped before the j^{th} round, for only the first three rounds of the simulations because very few audits, $(.1)^{j-1} \cdot (10^4)$ on average, make it to the j^{th} round. The proportions 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, when the underlying election is as announced. We see that, especially in earlier rounds for which the values are more representative of true audit behavior, our round size predictions are accurate. In particular, the average across all margins is just above $.9 = 90\%$ for all three rounds.

We also study the proportion of audits that stopped when the underlying election was a tie. This proportion should approach a value less than the risk limit, 0.1, as more audits are performed.

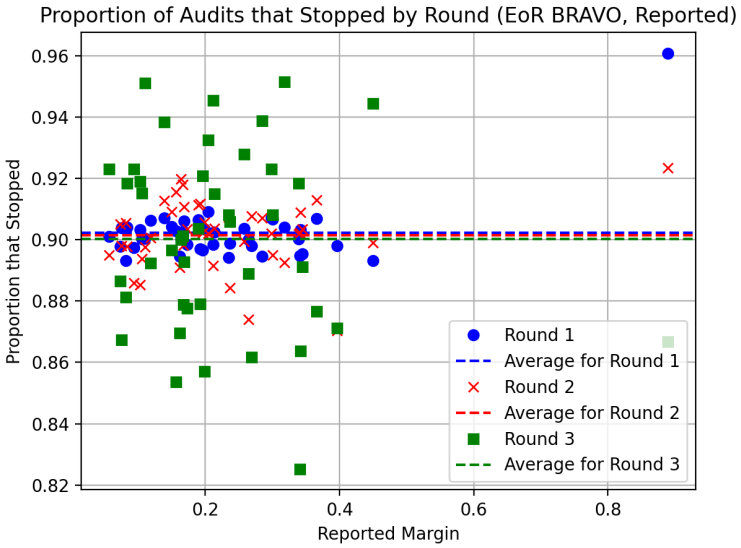

Fig. 1. This plot shows, for each state margin, when the underlying election is as announced, the number of EoR Bravo audits that stopped in the j^{th} round, as a fraction of all EoR Bravo audits which had not yet stopped before the j^{th} round for $j = 1, 2, 3$.

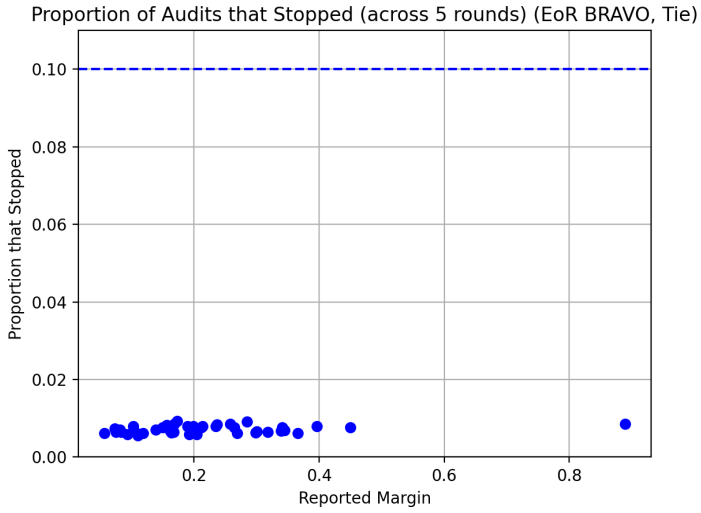

Fig. 2. This plot shows, for each state margin, the fraction of EoR Bravo audits that stopped in any of the 5 rounds when the underlying election was a tie.

We observe (see Figure 2) that the risk of EoR Bravo is roughly an order of magnitude less than the risk limit. These results are as expected, because EoR Bravo is known to be too conservative [17].

5.2 Selection-Ordered Bravo

For the SO Bravo simulations, our software estimated and used for each round the round size that would achieve a 90% probability of stopping.

In Figure 3, we again display proportions of SO Bravo audits that stopped in the j^{th} round to all audits which had not yet stopped before the j^{th} round for only the first three rounds. The proportions shown in Figure 3, like those in 1, give an estimate of the true probability of an SO Bravo audit stopping in the j^{th} round, given that it has not already stopped in a previous round. In 3, we see that our round size predictions are relatively accurate, all three rounds being near 90%.

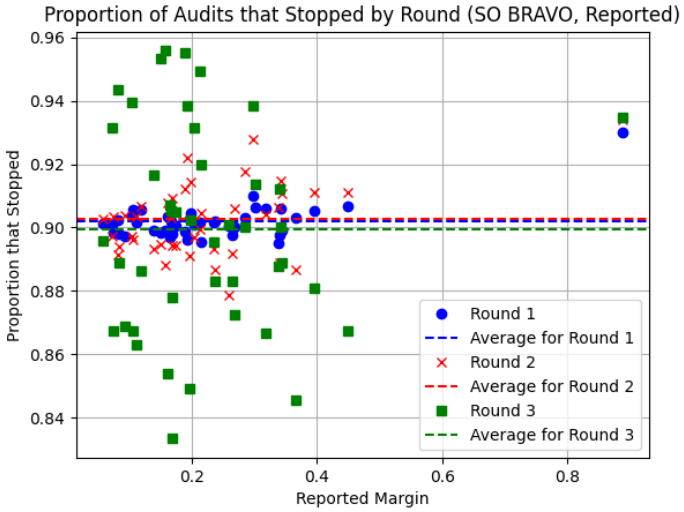

Fig. 3. This plot shows, for each state margin, when the underlying election is as announced, the number of SO Bravo audits that stopped in the j^{th} round, as a fraction of all EoR Bravo audits which had not yet stopped before the j^{th} round for $j = 1, 2, 3$.

As before, we will next consider the proportion of audits that stopped with an underlying tie over all five rounds. This proportion, for a risk-limiting audit, should approach a value less than the risk limit, 0.1, as more audits are performed.

In Figure 4 we show only the results for the 13 states whose simulations with an underlying tie have finished running. To estimate the next round size that

achieves a desired stopping probability, the SO Bravo software generates the probability distribution on the number of ballots in the sample ballot by ballot (see [17]) since the stopping condition needs to be evaluated for each individual ballot drawn. because the underlying tied election causes audits to move on to larger rounds, the simulations are computationally expensive. SO Bravo is proven to be a Risk-Limiting Audit, and we observe in Figure 4, that the risk of SO Bravo is much nearer the risk limit than that of EoR Bravo, as expected.

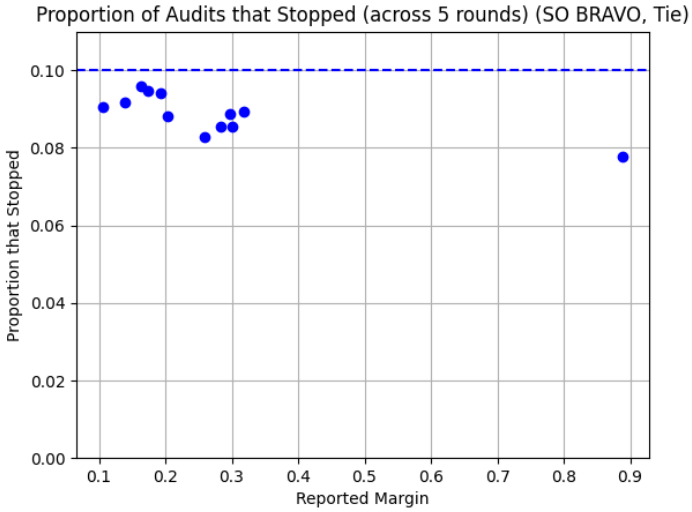

Fig. 4. This plot shows, for each state margin, the fraction of SO Bravo audits that stopped in any of the 5 rounds when the underlying election was a tie.

5.3 Minerva Simulations

The proof that Minerva is an RLA [17] assumes that the round schedule is predetermined. 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 use a first round size for 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 for MINERVA.

As with the preceding simulations, we ran $10,000 = 10^4$ trials per state for both an underlying election as reported and an underlying tied election.

Figure 5 and Figure 6 show that the first round size estimates were fairly accurate for multipliers of 10. and 1.5 respectively, with first round stopping probabilities being very close to 90%. For subsequent rounds, the multipliers of 1.0 and 1.5 respectively did not consistently achieve 90% stopping probability, as they were not accurate estimates of 90% stopping probabilities. Note that we chose a simple multiplier for future rounds, but one could make more accurate round size estimates before the audit begins.

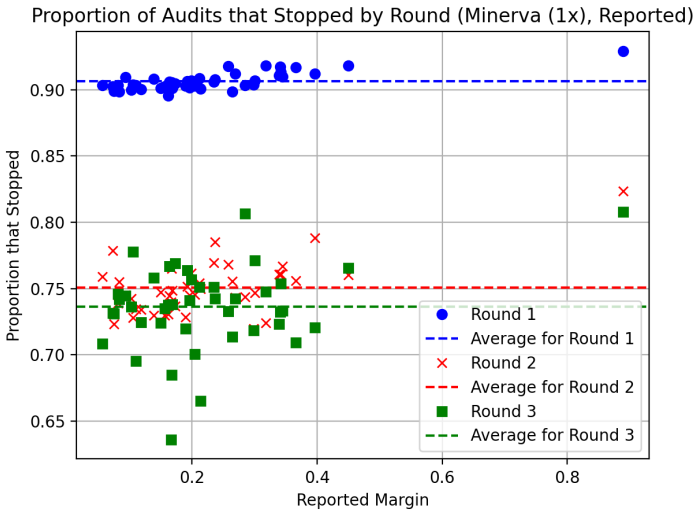

Fig. 5. This plot shows, for each state margin, when the underlying election is as announced, the number of MINERVA audits that stopped in the j^{th} round, as a fraction of all MINERVA audits which had not yet stopped before the j^{th} round for $j = 1, 2, 3$ and round size multiple of 1.0.

Figures 7 and Figure 8 show that fewer than 0.1 of the audits stopped when the underlying election was a tie, for round multiples of 1.0 and 1.5 respectively, as would be expected for an RLA with risk limit 0.1. 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.

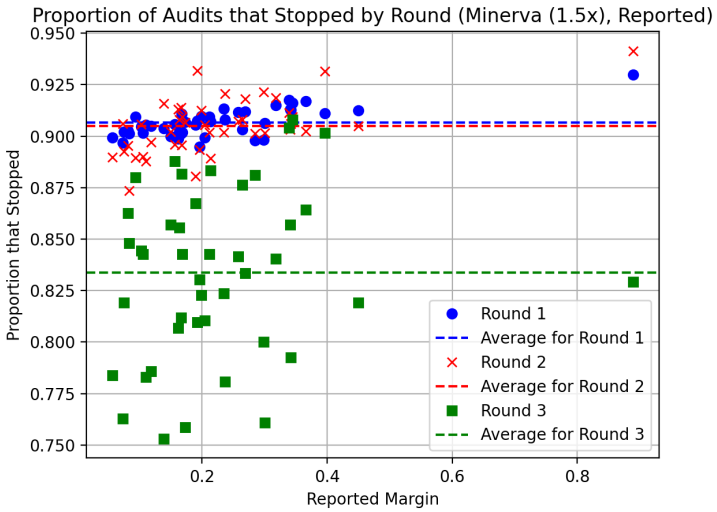

Fig. 6. This plot shows, for each state margin, when the underlying election is as announced, the number of MINERVA audits that stopped in the j^{th} round, as a fraction of all MINERVA audits which had not yet stopped before the j^{th} round for $j = 1, 2, 3$ and round size multiple of 1.5.

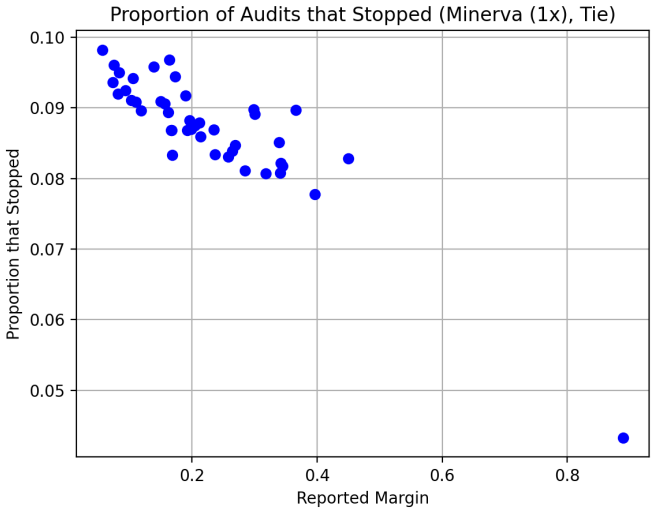

Fig. 7. This plot shows, for each state margin, the fraction of MINERVA audits with a round size multiple of 1.0 that stopped in any of the 5 rounds when the underlying election was a tie.

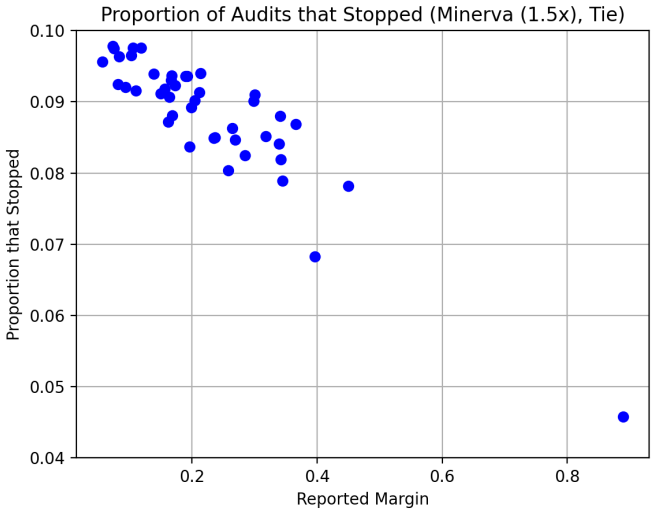

Fig. 8. This plot shows, for each state margin, the fraction of MINERVA audits with a round size multiple of 1.5 that stopped in any of the 5 rounds when the underlying election was a tie.

6 Conclusions and Future Work

References

1. Anonymized: Athena - risk limiting audit (round-by-round), https://github.com/xxxx
2. Anonymized: brla_explore, https://github.com/xxxx
3. Anonymized: R2B2 Basics, https://github.com/xxxx
4. Anonymized: R2B2 Comparing Audits with Graphs, https://github.com/xxxx
5. Anonymized: The R2B2 (Round-by-Round, Ballot-by-Ballot) library, https://github.com/xxxx
6. Blom, M.L., Stuckey, P.J., Teague, V.J.: Ballot-polling risk limiting audits for IRV elections. In: Krimmer, R., Volkamer, M., Cortier, V., Goré, R., Hapsara, M., Serdült, U., Duenas-Cid, D. (eds.) Electronic Voting - Third International Joint Conference, E-Vote-ID 2018, Bregenz, Austria, October 2-5, 2018, Proceedings. Lecture Notes in Computer Science, vol. 11143, pp. 17–34. Springer (2018). https://doi.org/10.1007/978-3-030-00419-4_2, `https://doi.org/10.1007/978-3-030-00419-4_2`
7. Huang, Z., Rivest, R.L., Stark, P.B., Teague, V.J., Vukcevic, D.: A unified evaluation of two-candidate ballot-polling election auditing methods. In: Krimmer, R., Volkamer, M., Beckert, B., Küsters, R., Kulyk, O., Duenas-Cid, D., Solvak, M. (eds.) Electronic Voting - 5th International Joint Conference, E-Vote-ID 2020, Bregenz, Austria, October 6-9, 2020, Proceedings. Lecture Notes in Computer Science, vol. 12455, pp. 112–128. Springer (2020). https://doi.org/10.1007/978-3-030-60347-2_8, `https://doi.org/10.1007/978-3-030-60347-2_8`

8. Lindeman, M., Stark, P.B., Yates, V.S.: BRAVO: Ballot-polling risk-limiting audits to verify outcomes. In: EVT/WOTE (2012)
9. McLaughlin, K., Stark, P.B.: Simulations of risk-limiting audit techniques and the effects of reducing batch size on the 2008 california house of representatives elections (2010), https://www.stat.berkeley.edu/users/vigre/undergrad/reports/McLaughlin_Stark.pdf
10. McLaughlin, K., Stark, P.B.: Workload estimates for risk-limiting audits of large contests (2011), https://www.stat.berkeley.edu/~stark/Preprints/workload11.pdf
11. Rivest, R.L., Shen, E.: A Bayesian method for auditing elections. In: EVT/WOTE (2012)
12. Stark, P.B.: Simulating a ballot-polling audit with cards and dice. In: Multidisciplinary Conference on Election Auditing, MIT (december 2018), http://electionlab.mit.edu/sites/default/files/2018-12/eas-ballotpollingsimulation.pdf
13. Vora, P.L.: Risk-limiting Bayesian polling audits for two candidate elections. CoRR **abs/1902.00999** (2019), http://arxiv.org/abs/1902.00999
14. VotingWorks: Arlo, https://voting.works/risk-limiting-audits/
15. Wald, A.: Sequential tests of statistical hypotheses. The Annals of Mathematical Statistics **16**(2), 117–186 (1945)
16. Zagórski, F., McClearn, G., Morin, S., McBurnett, N., Vora, P.L.: The Athena class of risk-limiting ballot polling audits. CoRR **abs/2008.02315** (2020), https://arxiv.org/abs/2008.02315
17. Zagórski, F., McClearn, G., Morin, S., McBurnett, N., Vora, P.L.: Minerva– an efficient risk-limiting ballot polling audit. In: 30th USENIX Security Symposium (USENIX Security 21). pp. 3059–3076. USENIX Association (Aug 2021), https://www.usenix.org/conference/usenixsecurity21/presentation/zagorski