Simulations of Ballot Polling Risk-Limiting Audits

Oliver Broadrick[1] Sarah Morin[1] Grant McClearn[2]
Neal McBurnett Poorvi L. Vora[1] Filip Zagórski[34]

[1]Department of Computer Science, The George Washington University (odbroadrick@gmail.com)

[2]Department of Computer Science, Stanford University (grantmcc@stanford.edu)

[3]Wroclaw University of Science and Technology (filip.zagorski@gmail.com)

[4]Votifica

April 12, 2022

Outline

- Risk-Limiting Audits
  - Bravo and Minerva
-

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 (SO) Bravo, and End-of-Round (EoR) Bravo. Bravo is the most commonly used ballot polling RLA and requires the smallest expected number of ballots when ballots are drawn one at a time and the (true) underlying election is as announced. In real audits, multiple ballots are drawn at a time, and Bravo is implemented as SO Bravo or EoR Bravo. Minerva is a recently proposed ballot polling RLA that requires fewer ballots than either implementation of Bravo in a first round with stopping probability 0.9 but requires a predetermined round schedule. It is an open question how these audits compare over multiple rounds and for lower stopping probabilities. Our simulations use stopping probabilities of 0.9 and 0.25. The results are consistent with predictions of the R2B2 open-source library for ballot polling audits. We observe that both Bravo audits are more conservative than Minerva, which stops with fewer ballots, for both first round stopping probabilities. However, the advantage of using Minerva decreases considerably for the smaller first

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) [**?**]. An RLA is an audit which guarantees that the probability of concluding that an election outcome is correct, given that it is not, is below a pre-determined value known as the risk limit of the audit, independent of the true, unknown vote distribution of the underlying election. Over a dozen states in the US have seriously explored the use of RLAs—some have pilot programs, some allow RLAs to satisfy a general audit requirement and some have RLAs in statute.

This paper provides insight into the main approaches to ballot polling RLAs, the Bravo audit [**?**], and the newer Minerva [**?**]

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, for two chosen probabilities of stopping (one high: 90%; the other low: 25%) if the election is as announced. This paper focuses on ballot-polling RLAs, which require a large number of ballots relative to comparison RLAs but do not rely on any special features of the election technology. Since comparison RLAs are not always feasible, ballot-polling audits remain an important resource and have been used in a number of US state pilots (California, Georgia, Indiana, Michigan, Ohio, Pennsylvania and elsewhere). In the general ballot-polling 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. Zagórski *et al.* [**?**] prove that

MINERVA is a risk-limiting audit and requires fewer ballots to be sampled than EoR BRAVO when an audit is performed in rounds, the two audits have the same pre-determined (before any ballots are drawn) round schedule and the underlying election is as announced[1] . They also present first-round simulations which show that MINERVA draws fewer ballots than SO BRAVO in the first round for first round sizes with a large probability of stopping when the (true) underlying election is as announced.
There are no results, either theoretical or based on simulations, regarding the number of ballots drawn over multiple rounds in a MINERVA audit with a pre-determined schedule. Because BRAVO does not need to work on a pre-determined round schedule, it can optimize the size of the next round based on the sample drawn so far. Thus an open question is whether the constraint of a predetermined round schedule limits the efficacy of MINERVA in future rounds, and there is no literature comparing the number of ballots drawn by MINERVA and SO BRAVO over multiple rounds. Note that the Average Sample Number (ASN) computations for BRAVO [**?**] apply only for B2 audits and are especially misleading

as estimates of the number of ballots drawn over multiple rounds when first round sizes are large.

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

For a two candidate plurality contest with a risk limit of 10%, we observe the following about the total number of ballots drawn over five rounds:

1. Even when the first round stopping probability is as small as 0.25, the number of ballots required for MINERVA is smaller than that required by SO BRAVO and EoR BRAVO. However, the improvement is considerably smaller than that when the stopping probability is 0.9.
    - The number of ballots required by SO BRAVO for a first round stopping probability of 0.9 is about a third more than that required by MINERVA. On the other hand, for a first round stopping probability of 0.25, it requires only about a tenth more ballots than does MINERVA.
    - The number of ballots required by EoR BRAVO for a first

round stopping probability of 0.9 is about twice those required by MINERVA. On the other hand, for a first round stopping probability of 0.25, it requires only about a fourth to a half more ballots (depending on margin) than does MINERVA.

2. For a first round stopping probability of 0.9, when consequent MINERVA rounds are the same size (multiplying factor 1), consequent conditional stopping probabilities are about 0.75 and 0.74 respectively for rounds two and three. When the multiplying factor is 1.5, the conditional stopping probabilities for rounds two and three are 0.91 and 0.83 respectively. Both our simulator and the code estimating probabilities and round sizes are flexible enough to enable the study of various predetermined round schedules.

Section 2 describes related work. The experiments we performed are described in section 3, and sections 4 and 5 present our results. Section 6 has our conclusions.

The BRAVO audit [**?**] 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) [**?**], 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 large numbers of ballot, called rounds, before making stopping decisions because sequentially sampling individual ballots has significant overhead (unsealing storage boxes and searching for individual ballots). It is possible to apply BRAVO to the sequence of ballots in a round 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 ballots that were drawn earlier in the sample.

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 estimated as implemented in [**?**]; this method is due to Mark Lindeman and uses quadratic approximations. A later method for stopping probability estimates presented by Zagórski *et al.* [**?**, **?**] uses a similar

technique for narrow margins and a separate algorithm for wider margins, the results of which match simulation results reported by Lindeman *et al.* [**?**, Table 1].

The MINERVA audit [**?**, **?**] 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 inform 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 are smaller than those for 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.

Ballot polling audit simulations have been used to familiarize election officials and the public with the approach [**?**]. McLaughlin

and Stark [**?**, **?**] 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 [**?**]. Huang *et al.* present a framework generalizing a number of ballot polling audits and compare their performance (round sizes and stopping probabilities) using simulations [**?**]. This work was prior to the development of MINERVA, and focuses on the comparison between Bayesian audits [**?**] and BRAVO, essentially studying the impact of the prior of the Bayesian RLA. Some workload measurements have been made [**?**]. While total ballots sampled can give naive workload estimates [**?**], Bernhard presents a more complex workload estimation model [**?**].

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. Note that sampling without replacement is more efficient for large sampling fractions, but MINERVA has not been extended for sampling without replacement. We first present relevant

definitions.

Definition

An audit $\mathcal{A}$ takes a sample of ballots X as input and gives as output either (1) *Correct*: the audit is complete, or (2) *Uncertain*: continue the audit.

All of the audits discussed in this paper are modeled as binary hypothesis tests. Under the alternative hypothesis, H_a, the announced outcome is correct. In particular, the true underlying ballot distribution is given by the announced ballot tallies. Under the null hypothesis, H_0, a tie is the correct outcome [2]. The maximum risk of an audit is the probability that an audit stops, given that the underlying election is a tie [**?**]. Note that an audit $\mathcal{A}$ includes all audit parameters (maximum risk, round sizes, etc.).

Definition (Risk)

The maximum risk R of audit $\mathcal{A}$ with sample $X \in \{0,1\}^*$ drawn from the true underlying distribution of ballots is $R(\mathcal{A}) = \Pr[\mathcal{A}(X) = \textit{Correct} \mid H_0]$.

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

Definition (Risk Limiting Audit (α-RLA))

An audit $\mathcal{A}$ is a Risk Limiting Audit with risk limit α iff $R(\mathcal{A}) \leq \alpha$.

We present measures of stopping probability in the j^{th} round of the audit, given that the underlying election is as announced.

Definition (Stopping Probability)

The stopping probability S_j of an audit $\mathcal{A}$ in round j is

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

Experimentally, using our simulations, S_j would be estimated by the fraction of audits that stop in round j. Note that $\sum_j S_j(\mathcal{A}) = 1$.
We can also consider the cumulative stopping probability:

Definition (Cumulative Stopping Probability)

The cumulative stopping probability C_j of an audit $\mathcal{A}$ in round j is $C_j(\mathcal{A}) = \sum_{i=1}^{j} S_j$

Experimentally, using our simulations, C_j would be estimated by the fraction of audits that stop in or before round j.
Finally, we are also interested in the probability that an audit will

stop in round j given that it did not stop earlier:

Definition (Conditional Stopping Probability)

The conditional stopping probability of an audit $\mathcal{A}$ in round j is

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

Experimentally, using our simulations, χ_j would be estimated by the ratio of the audits that stop in round j to those that "entered" round j, i.e. those that did not stop before round j.
We simulated audits for a risk limit of 10% (as in [**?**] and [**?**]) using margins from the 2020 US Presidential election, limiting ourselves to pairwise margins for the two main candidates of 0.05 or larger. Note that both BRAVO and MINERVA can be extended for multiple-candidate, multiple-winner plurality contests by performing pairwise tests between the winners and the losers [**?**, **?**]. Therefore, the two candidate plurality contest is a general case, and these simulations provide insight for multiple-candidate and multiple-winner contests too. 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 (H_a), and an additional $10,000 = 10^4$ audits assuming the underlying election was a tie (H_0).

We ran simulations for: (a) 0.9 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 and (b) .25 probability of stopping in the first round which is more favorable to BRAVO. We ran our simulations for up to five rounds.

For rounds after the first, we chose the round schedule as follows. For both versions of BRAVO, we chose a single round schedule: each round size has the same conditional stopping probability as the first one. As the proof of the risk-limiting property of MINERVA assumes that its round schedule is determined before any ballots are drawn, we could not use this approach for MINERVA round sizes. Instead, we chose to compare two fixed round schedules for MINERVA: one where the additional number of ballots drawn in a round is the same as in the previous round

(multiplying factor of 1.0) and the second where the multiplying factor is 1.5. 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 version is integrated into *Arlo*, and the multiplying factor was chosen as it roughly ensures a 0.9 conditional stopping probability in the second round for a first round stopping probability of 0.9. We used the R2B2 library [**?**], which provides a framework for the exploration of round-by-round and ballot-by-ballot RLAs. It 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, 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 as announced). When there are more than two candidates, 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.

For both SO and EoR BRAVO simulations, our software estimated round sizes that would give $\chi_j(\mathcal{A}) = 0.9$ and used those for the simulations. In Figure 1, we display the proportion of EoR BRAVO audits that stopped in the j^{th} round to all audits which had not stopped before the j^{th} round, for $j = 1, 2, 3$. Though we carried out the simulations for 5 rounds we show 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 for $j \geq 4$. In Figure 2, we display the same proportions for SO BRAVO audits. In both cases, these proportions are estimates of the true value of $\chi_j(\mathcal{A})$ for $j = 1, 2, 3$ as a function of margin. We see that, especially in earlier rounds for which the values are more representative of true audit behavior because fewer simulated audits have stopped, our round size predictions are accurate (the proportions are close to 0.9).

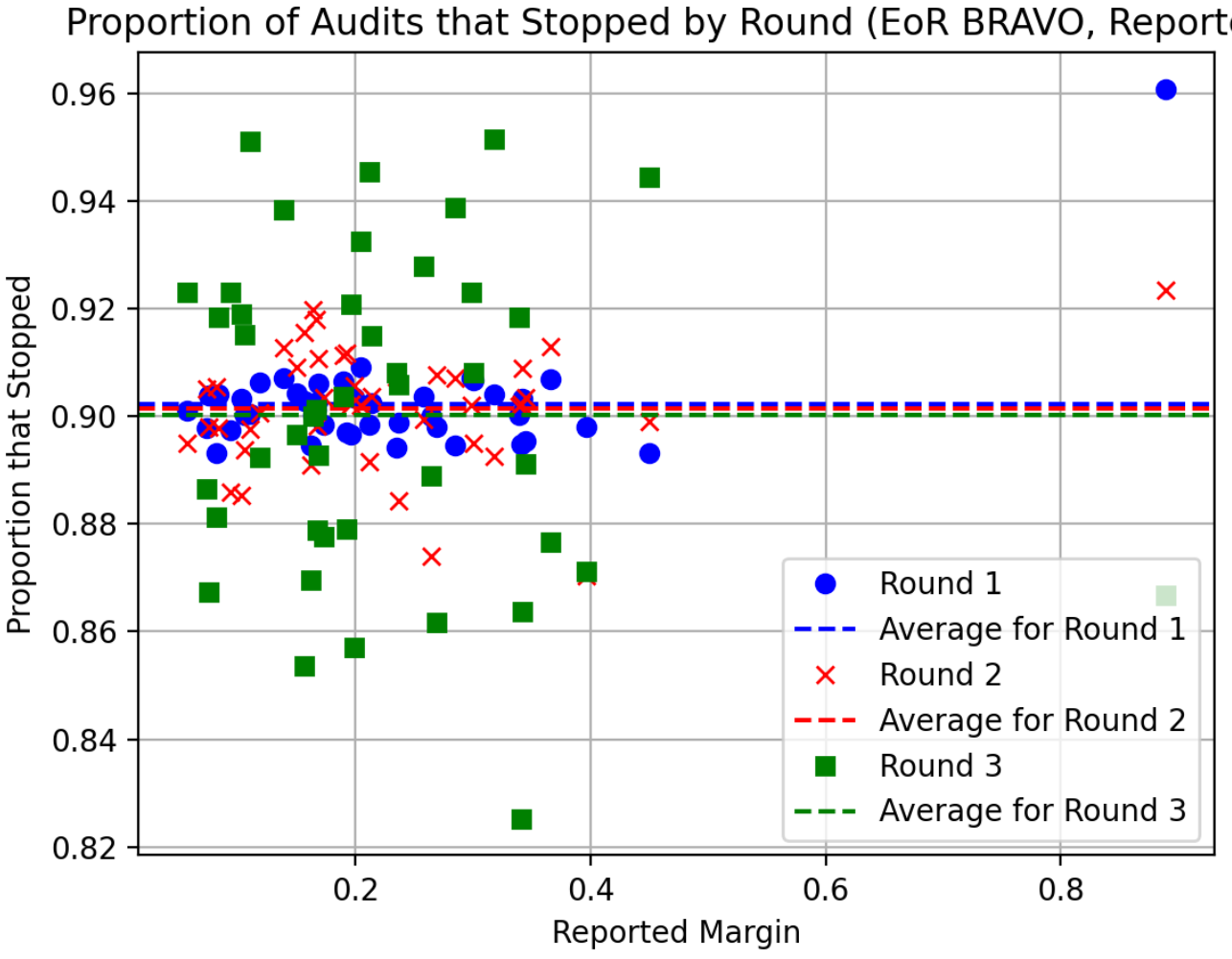

Figure: This plot shows, for each state margin, when the underlying election is as announced, the number of EoR BRAVO audits that stopped in the j^{th} round, as a fraction of all EoR BRAVO audits which had not yet stopped before the j^{th} round for $j = 1, 2, 3$ and $S_1 = 0.9$.

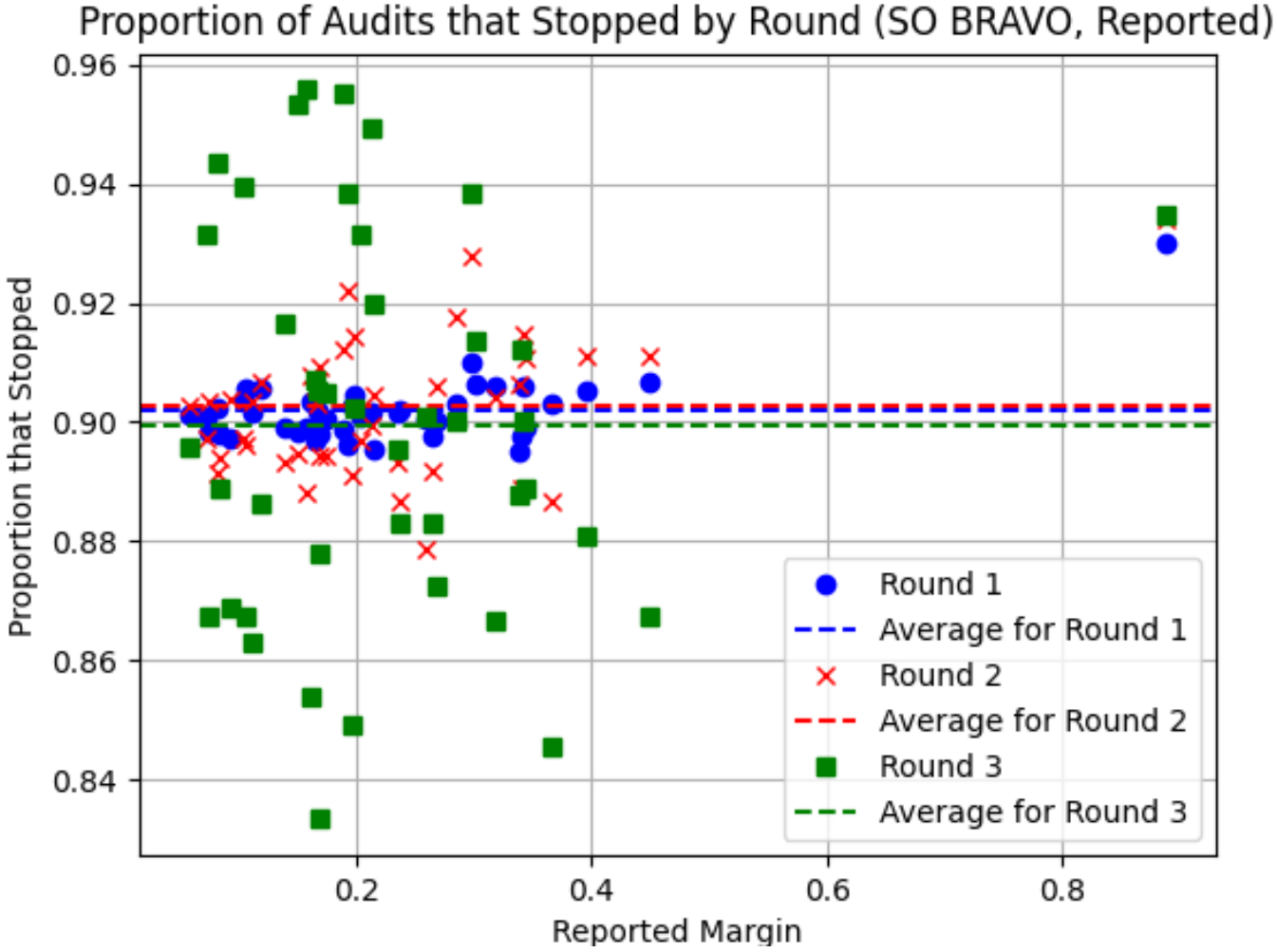

Figure: This plot shows, for each state margin, when the underlying election is as announced, the number of SO BRAVO audits that stopped in the j^{th} round, as a fraction of all SO BRAVO audits which had not yet stopped before the j^{th} round for $j = 1, 2, 3$ and $S_1 = 0.9$.

Figure 3 and Figure 4 show the same proportions for MINERVA

round multipliers of 1.0 and 1.5 respectively. We see that the first round size estimates were fairly accurate, with first round stopping probabilities being very close to .9. For subsequent rounds, the multipliers of 1.0 achieved smaller stopping probabilities, as it was not chosen so as to obtain $\chi_j(\mathcal{A}) = 0.9$. The 1.5 multiplier is a good estimate for $j = 2$, but the stopping probability for $j = 3$ is slightly smaller than 0.9. Note that we chose a simple multiplier for future rounds, but one could make more accurate round size estimates before the audit begins.

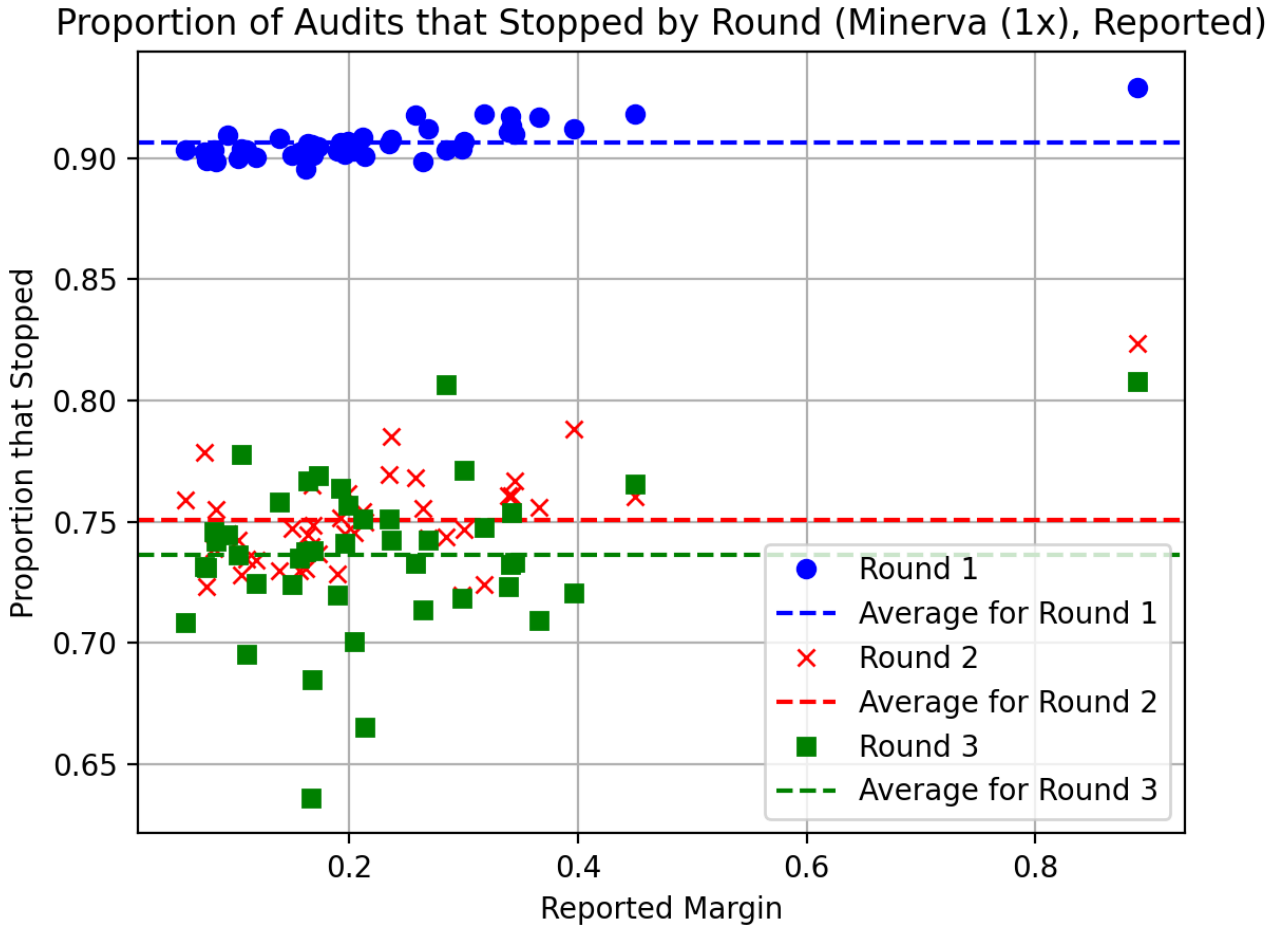

Figure: This plot shows, for each state margin, when the underlying election is as announced, the number of MINERVA audits that stopped in the j^{th} round, as a fraction of all MINERVA audits which had not yet stopped before the j^{th} round for $j = 1, 2, 3$, round size multiple of 1.0 and $S_1 = 0.9$.

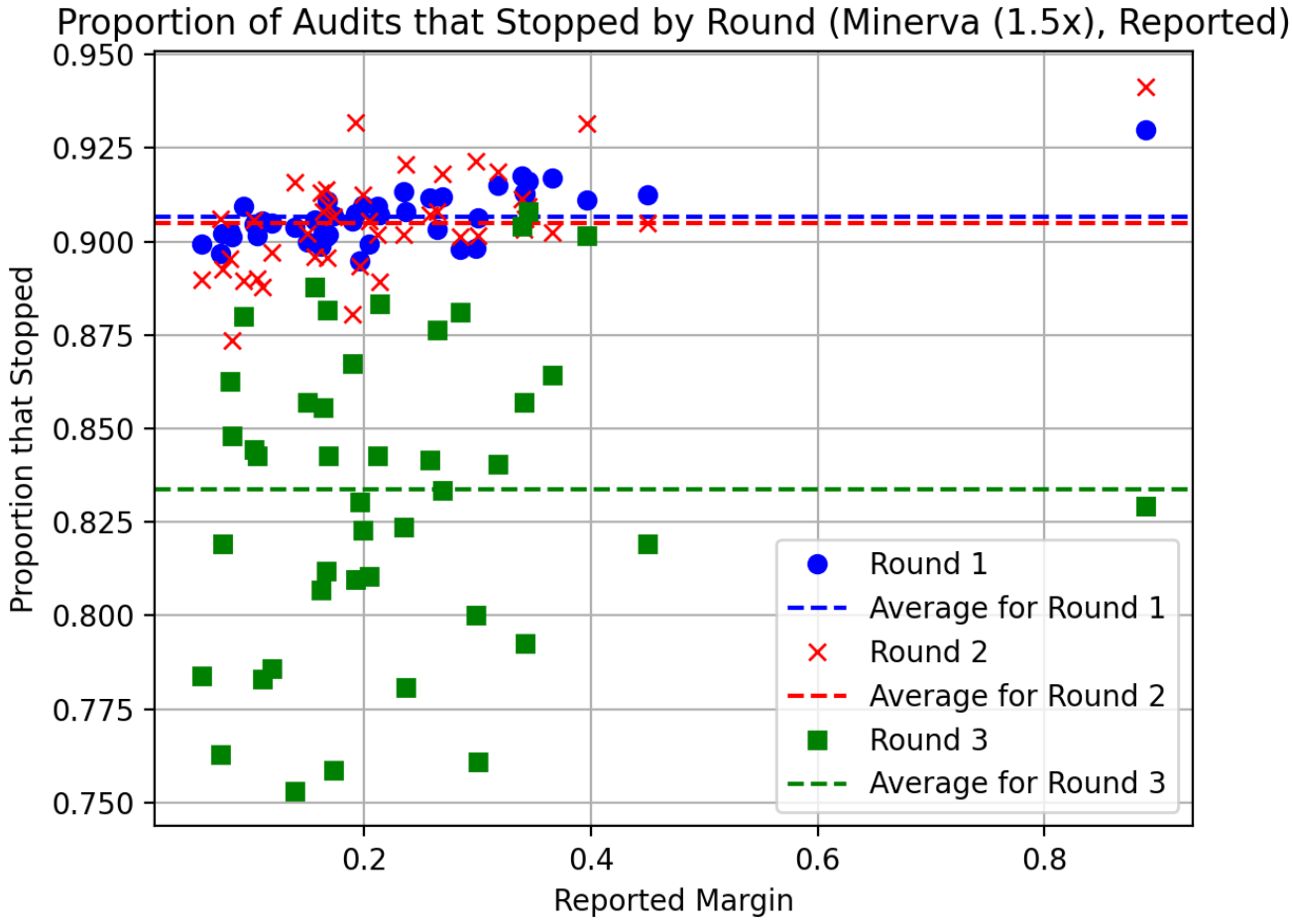

Figure: This plot shows, for each state margin, when the underlying election is as announced, the number of MINERVA audits that stopped in the j^{th} round, as a fraction of all MINERVA audits which had not yet stopped before the j^{th} round for $j = 1, 2, 3$, round size multiple of 1.5 and $S_1 = 0.9$.

Finally, we can perform a similar study for $S_1 = 0.25$. See Figure 5 for an example, MINERVA with round mutiplier 1.5.

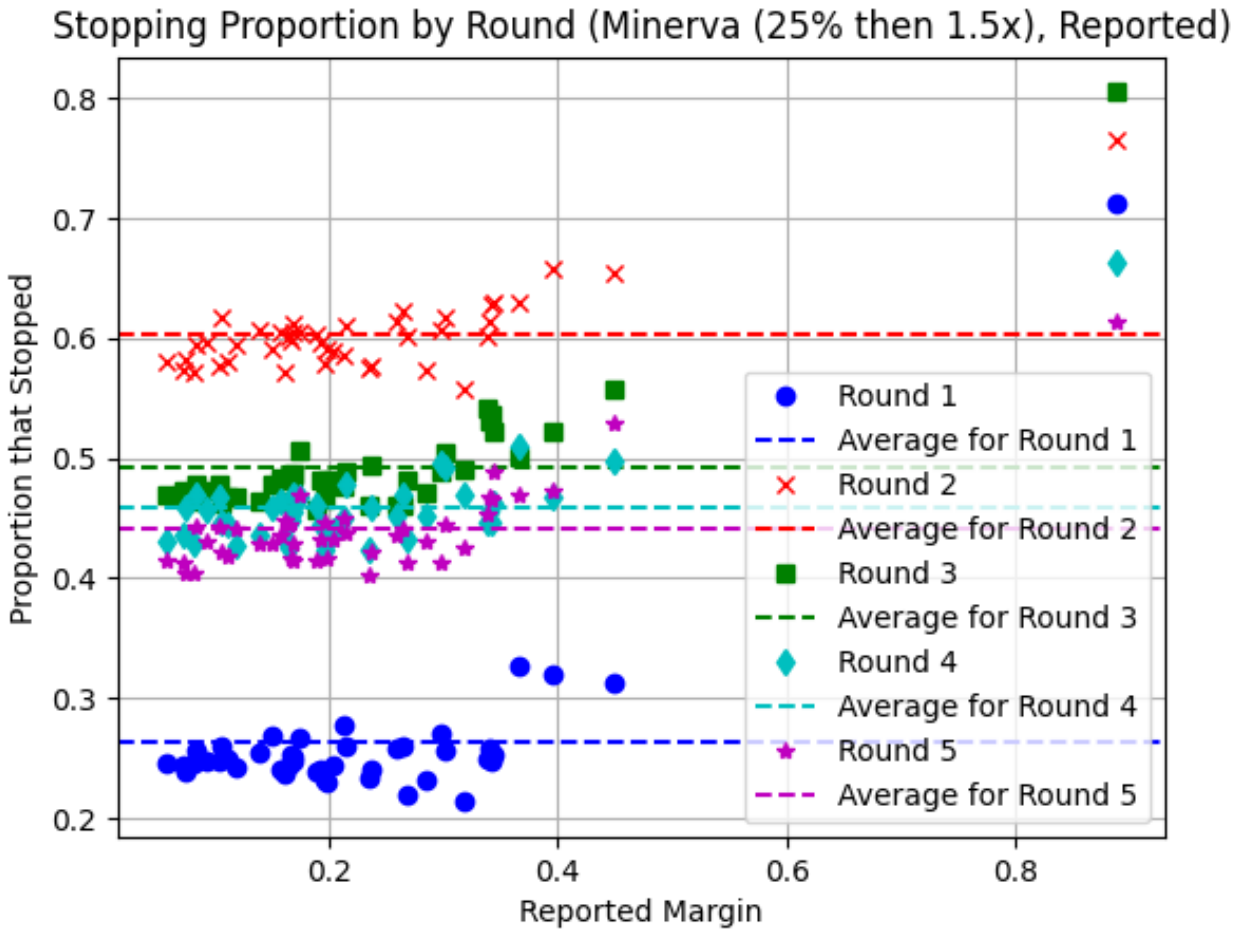

Figure: This plot shows, for each state margin, when the underlying election is as announced, the number of MINERVA audits that stopped in the j^{th} round, as a fraction of all MINERVA audits which had not yet

stopped before the j^{th} round for $j = 1, 2, 3$, round size multiple of 1.5 and $S_1 = 0.25$.

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, 10%, as more audits are performed.

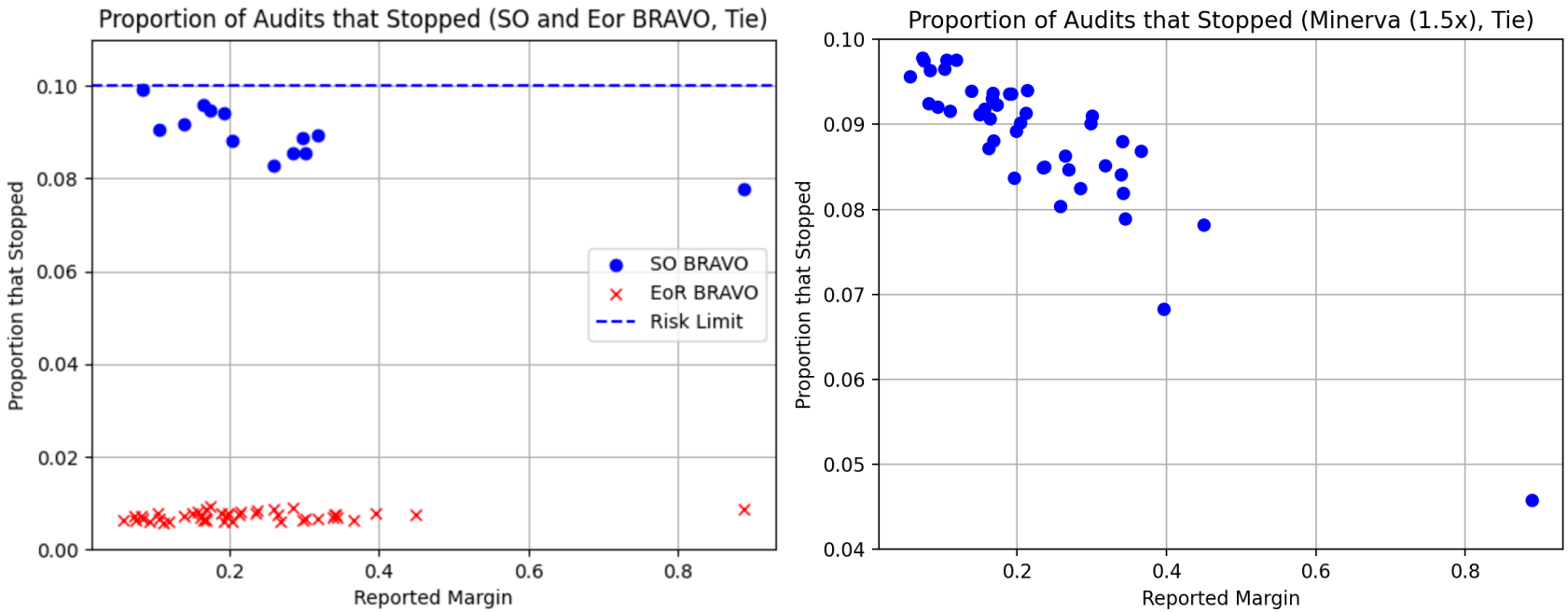

Figure: The left hand plot shows the fraction of EoR BRAVO audits (all states with margins at least 0.05) and SO BRAVO audits (the 13 states for which our simulations are complete so far) that stopped in any of the

5 rounds when the underlying election was a tie. The right hand plot, for each state margin, shows 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.

We observe 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 [**?**].
In Figure 6 we show only the results for the 13 states for which our simulations with an underlying tied election have completed. 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 [**?**]) 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 6, that the risk of SO BRAVO is much nearer the risk limit than that of EoR BRAVO, as expected.

Figure 6 shows that fewer than 0.1 of the audits stopped when the underlying election was a tie, for round multiples 1.5, as would be expected for an RLA with risk limit 10%. 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. The plot for round multiple 1.0 is very similar.

In this section we present our data on the expected number of ballots drawn as the number of rounds increases, and on the fraction of audits that stop (an estimate of cumulative stopping probability, C_j) for the states of Texas, Missouri and Massachusetts, with margins of 0.057, 0.157 and 0.342 respectively. Interestingly, we observe that MINERVA has an advantage for a first round size with stopping probability $S_1 = 0.25$, but it is not as large as that for $S_1 = 0.9$. On all our plots we mark ASN, the Average Sample Number for B2 BRAVO for context. Notice that, in all the plots, both instances of MINERVA show a higher probability of completion than does either BRAVO audit when the average number of ballots drawn is ASN.

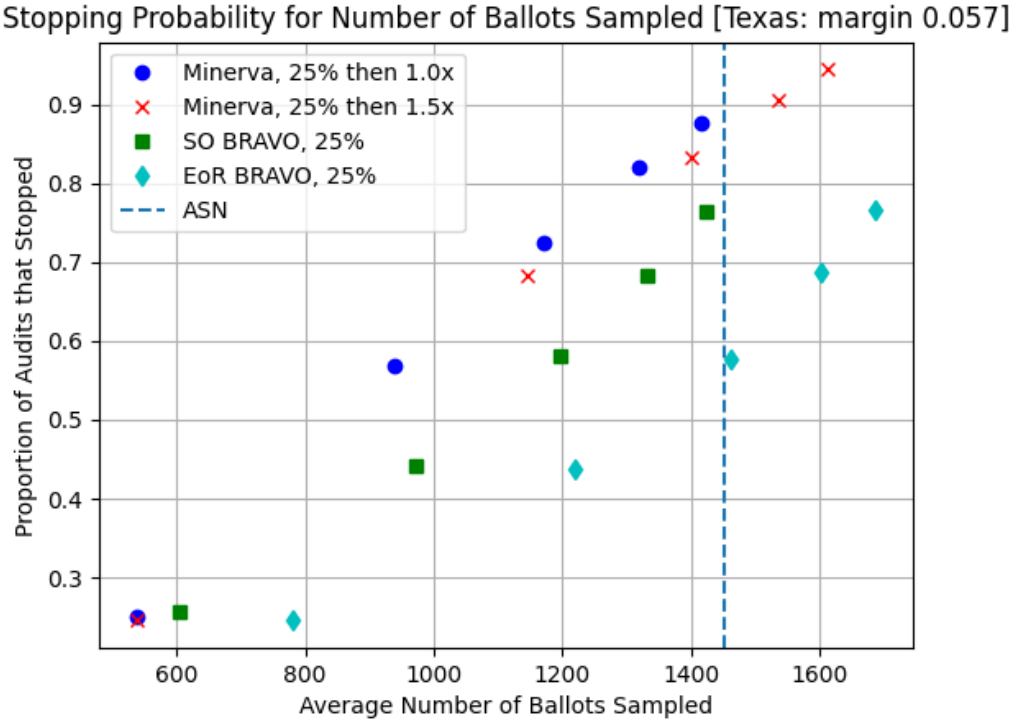

Figure: This plot shows the cumulative fraction of audits that stopped as a function of average number of sampled ballots for all four audits we studied, for the state of Texas, margin 0.057, and first round stopping probability $S_1 = 0.25$.

We observe that the behavior of both MINERVA audits is similar, and that the plot for SO BRAVO is to the right of (more ballots) and below (lower probability of stopping) those for MINERVA, even for a stopping probability as low as 0.25. We observe that the plot for EoR BRAVO shows the worst performance, which is not

surprising. We observe similar behavior across margins (see Figures 8 and 9), though the improvement due to MINERVA reduces as margins get larger. We see also that the improvement due to using MINERVA is not as large as that seen for $S_1 = 0.9$ (see Figure 10). For $S_1 = 0.25$, the ratio of first round size of EoR BRAVO to MINERVA is 1.45, 1.37, 1.23 for states Texas, Missouri and Massachusetts, and margins 0.057, 0.157 and 0.342 respectively. This may be compared to 2.03, 1.99 and 1.8 respectively for $S_1 = 0.9$. Similarly, for $S_1 = 0.25$, the ratio of first round size of SO BRAVO to MINERVA is 1.13, 1.08, 1.12 for states Texas, Missouri and Massachusetts, and margins 0.057, 0.157 and 0.342 respectively. This may be compared to 1.38, 1.38 and 1.30 respectively for $S_1 = 0.9$. Note that the effect of such improvements on workload depends greatly on the number of ballots being sampled. For example, a 20% reduction in sample size in Massachusetts might save election officials 10 ballots, whereas the same reduction in Texas could save thousands.

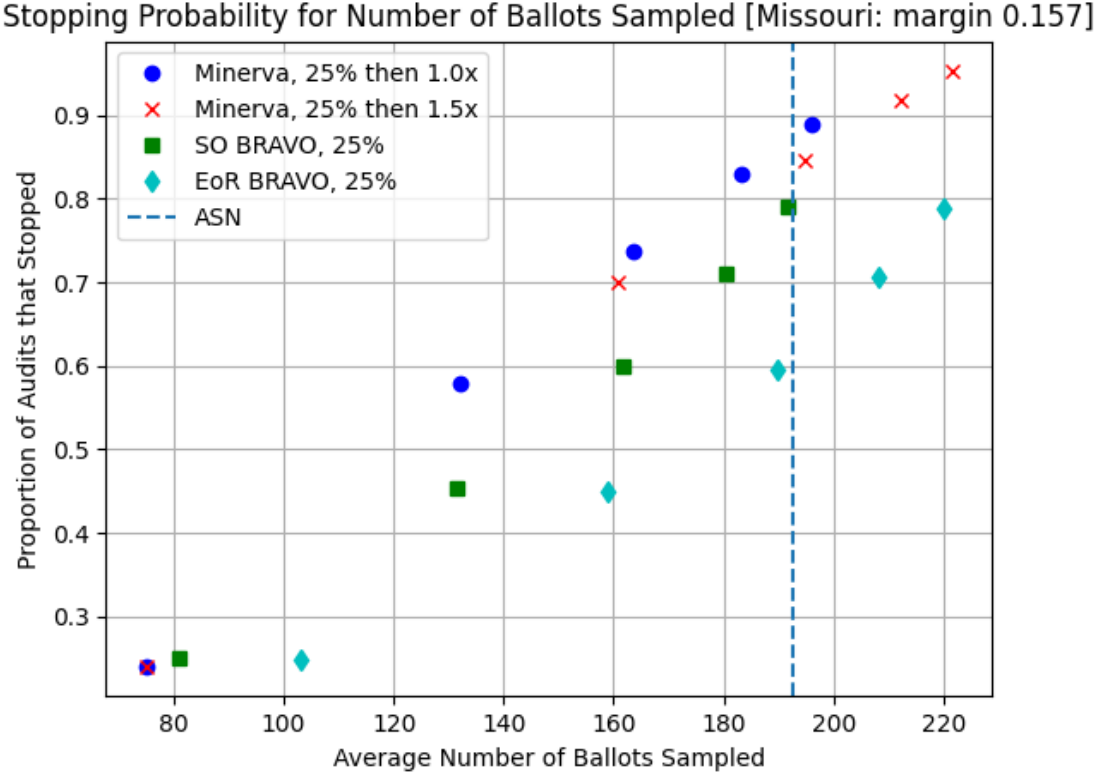

Figure: This plot shows the cumulative fraction of audits that stopped as a function of average number of sampled ballots for all four audits we studied, for the state of Missouri, margin 0.157, and first round stopping probability $S_1 = 0.25$.

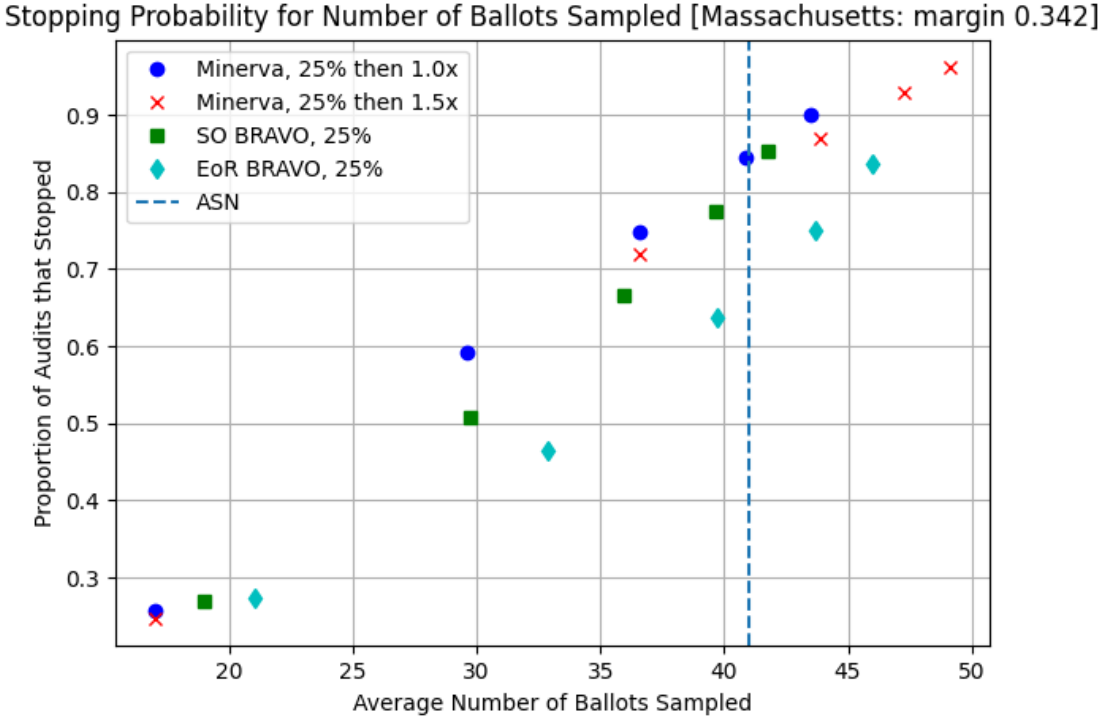

Figure: This plot shows the cumulative fraction of audits that stopped as a function of average number of sampled ballots for all four audits we studied, for the state of Massachusetts, margin 0.342, and first round stopping probability $S_1 = 0.25$.

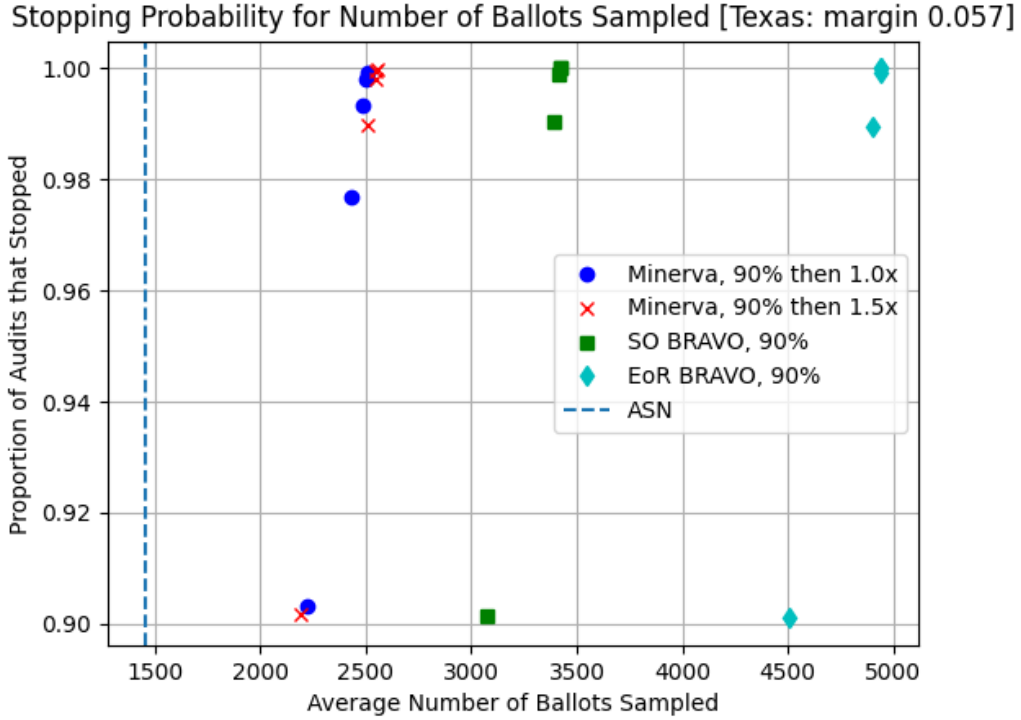

Figure: This plot shows the cumulative fraction of audits that stopped as a function of average number of sampled ballots for all four audits we studied, for the state of Texas, margin 0.057, and first round stopping probability $S_1 = 0.9$.

We describe the use of the R2B2 library and simulator to characterize the maximum risk, stopping probability and average number of ballots required across round schedules which may be specified through conditional stopping probabilities (as with the BRAVO audits) or pre-determined round sizes (as with MINERVA).

We use simulations to study the number of ballots drawn when the first round size is small (stopping probability of 0.25) and when it is large (stopping probability of 0.9) for a risk limit of 0.1. We observe that the advantage of using MINERVA is smaller for the smaller stopping probability of the first round, as would be expected. We also observe that MINERVA does still require fewer ballots all the way through five rounds.

A promising direction for future work would be a more detailed study of the impact of first round stopping probability and different round schedules on overall stopping probability and number of ballots for both MINERVA and BRAVO.