JETS

The USENIX Journal of Election Technology and Systems

Volume 3, Number 1 • December 2014

usenix
THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

JETS

The USENIX Journal of Election Technology and Systems

Volume 3, Number 1 • December 2014

JETS articles in this issue will be presented at the 2015 USENIX Journal of Election Technology and Systems Workshop (formerly EVT/WOTE).

ISBN 978-1-931971-195 ISSN 2328-2797

usenix
THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

USENIX Journal of Election Technology and Systems (JETS)
Volume 3, Number 1 • December 2014
www.usenix.org/jets/issues/0301

DIFFUSION OF VOTER RESPONSIBILITY: POTENTIAL FAILINGS IN E2E VOTER RECEIPT CHECKING

ESTER MOHER, Children's Hospital of Eastern Ontario Research Institute
JEREMY CLARK, Concordia University
ALEKSANDER ESSEX, Western University

End-to-end verifiable (E2E) voting systems provide voters with (privacy preserving) receipts of their ballots allowing them to check that their votes were correctly included in the final tally. A number of recent studies and field tests have examined the usability of ballot casting and receipt checking. Simply checking receipts, however, is not enough to provide strong assurance that the election outcome is correct; voters must also be counted on to report any discrepancies between their receipts and the official record when they occur. In this paper we designed and ran a study examining the frequency and conditions under which voters (a) check their receipts, and (b) report discrepancies when they occur. Participants were recruited online and were asked to vote in a survey on charitable giving. Similar to previous work, we found that the proportion of voters performing a receipt check was low. More importantly, within this group, we found that the proportion of voters reporting discrepancies was also low. We did however observe that the incidence of receipt checking was significantly higher when the election outcome was unanticipated or unexpected by voters. In the condition with an adverse election result we observed that, while 7.5% of voters checked receipts, only 0.5% filed a dispute when shown an incorrect receipt. With such low reporting rates, E2E voting systems will struggle to detect fraud with high confidence, especially in elections with narrow margins of victory. We posit, therefore, that improving the usability of the receipt check component in E2E systems is an important open problem.

1. INTRODUCTION

Relative to conventional ballot casting methods, end-to-end verifiable (E2E) voting systems can be designed to provide cryptographically strong security guarantees. Depending on the specifics of the system, the correctness of the election outcome can be stated unconditionally (*i.e.,* without requiring any cryptographic or computational assumptions), even in the face of a corrupt election authority. Nearly all E2E systems share a common design element: after casting her ballot, the voter retains a privacy-preserving record of her vote called a *receipt*. After the election, the election authority publishes a list of receipts it asserts to have collected. This list is cryptographically transformed into a corresponding tally, hiding the voter-vote association and ensuring that any manipulation of this process is detectable with overwhelming probability.

If the election authority misbehaves and publishes an incorrect receipt list, the votes associated with these receipts could be modified. As one of the practical limitations of the end-to-end verification paradigm, the cryptographic checks cannot detect incorrect receipts; the burden falls to the voters to check. In order to detect maliciously modified receipts, the security guarantees of E2E systems have typically rested on the tacit assumption that a sufficient proportion of voters will (i) check that their receipts match the list produced by the election authority, (ii) report any discrepancies, and (iii) be able to convince others of the validity of an honest dispute. The system must be able distinguish between a voter mistake or lie (*e.g.,* to cast doubt on the outcome), and a malicious election authority. Recent trial elections and user studies provide some data on the first property: between 4% and 54% of voters check their receipts depending on the election. Contemporary E2E systems like Scantegrity II [Carback et al. 2010] provide a mechanism to ensure the third property, sometimes called *dispute resolution* or *accountability*. To our knowledge, however, the second property has not been explicitly considered in the literature. It is often implicitly assumed any voter checking their receipt will always uncover and report any discrepancy found. There are, however, a variety of reasons that a voter might not legitimately file a dispute: a voter might check but not notice the discrepancy, notice but assume the fault is their own, acknowledge the error should be reported but fail to do so because its too much work, or conclude the responsibility of reporting rests with another party.

In this paper, we design a study to measure how many voters are willing to report errors relative to how many check their receipts. We also examine how verification rates change according to how expected or unexpected the outcome of the election is. We find verification is higher when the result is unexpected (as might be expected in a manipulated tally), but even when voters are provided an unexpected result and wrong receipt, only 0.5% of voters actually report the discrepancy (versus 7.5% who simply check). To illustrate the consequences of this drop-off, we apply a recent US Senate race with a slim margin of victory to a fictional scenario where enough receipts were manipulated to change the election outcome. We then show that the probability of the electorate detecting such an attack drops from 99.99% to 43.25% when we factor in the low rate at which voters actually report wrong receipts.

2. RELATED WORK

2.1 End-to-end verifiable (E2E) voting systems

The literature on end-to-end verifiable (E2E) voting is vast. Staring in 1996 with a variant of FOO at Princeton, a number of E2E systems have been developed and deployed in student elections including Evox, Punchscan, Bingo Voting, Helios, and Wombat [Clark, 2011]. In 2009, Scantegrity II was used in a governmental election in Takoma Park, USA, and Prêt à Voter will be deployed in a federal election in Victoria, Australia in 2014. Although these systems differ in many details, they share a common approach to voter verifiability. Voters mark their selections, and the receipt consists of a privacy-preserving obfuscation of the selections along with a unique identifier. Different vote obfuscation methods have been proposed: an encryption of the selection, the position of the selection in a shuffled list, or short alphanumeric codes (confirmation codes) for each selection. In our study, we utilize confirmation codes.

After the election, the voter may use their receipt identifier to lookup her receipt and confirm it was recorded correctly. A second verification step (which is not a focus of this work) is performed to verify that all receipts are correctly processed to produce a final tally. Any manipulation of the association between a receipt and its corresponding vote has an overwhelming probability of detection. However the probability of detecting manipulations of receipts themselves depends on how many voters check and, importantly, how many of these actually report the discrepancy.

2.2 Reporting

The literature has focused to a great extent on the proportion of individuals who check their vote as a proxy for the number of voters who will report potential voter fraud. Two theories, however, suggest that even when a voter checks their vote and observes that an error has occurred, they may be unlikely to report a problem.

The first issue relates to the psychological phenomenon of "diffusion of responsibility", or the bystander effect [Darley and Latane 1968; Latane and Darley 1970]. For example, when a crime is committed and many people witness it, each individual witness is less likely to call the police and report the crime that when few people witness the crime. That is, individuals do not take steps to act responsibly (by reporting the crime), because they assume someone else in the crowd will do it [Darley and Latane 1968]. We hypothesize similar phenomena may arise in E2E elections, *i.e.,* individuals who are faced with a potential voting error may assume that if others have also experienced the same problem, others will have reported the problem. The result is a potentially large proportion of erroneous votes going unreported.

A second issue relates to lack of expertise in understanding the purpose and process of E2E election verification. Voters who are unfamiliar with E2E systems may attribute an inconsistent receipt with their own error, as opposed to the error of the system. As a result, individuals may be less likely to report a problem than has been previously anticipated.

2.3 Voter Verification

The Punchscan system was deployed at U. Ottawa [Essex et al 2007] with a reported 54% of receipts queried, and no reported disputes. Helios was deployed at UC Louvain [Adida et al. 2009] with a reported 30% of receipts queried, with 7 voters filing a dispute. Scantegrity II was used in Takoma Park, MD [Carback et al. 2010] with a reported 4% of receipts queried, with 1 voter filing a dispute. In a comparative usability study of Helios, Scantegrity II and Prêt à Voter, a reported 43% and 38% of voters attempted receipt verification for Helios and Scantegrity II respectively [Acemyan et al. 2014].

In a questionnaire study, five prominent mental models of verifiability were identified [Olembo et al. 2013]: the first group references a belief that persons and/or processes are trustworthy, the second that verifiability is not possible or are of unsure it, the third that the presence of external observers can ensure integrity, the fourth that personal involvement (or the potential for it) can ensure integrity, and the final group that references auditing techniques. We do not study which groups our participants fall in, however our focus on the voters who do verify their vote likely correlate to the final group.

In later work, a study examines if textual prompts can increase a voter's intent to verify their vote [Olembo et al. 2014]. It is reported that prompts do increase intent with no measurable difference between the types of message (*e.g.,* communicating a risk vs. a social norms). Independently in this work, we chose to use a risk-based prompt to increase verification. We also measure actual verification rates as opposed to stated intentions to verify.

2.4 DRE Review Screen Verification

A user study of DRE voting machines reports that 37% of voters noticed when the review screen summarizing the voter's selections prior to casting the ballot contained manipulations (vote flips) [Everett 2007]. Similarly in our study, receipt information was manipulated (wrong code) when shown on a review page. Our study differs in a few regards that may impact the results: (i) review screens are mandatory, while receipt checking is opt-in, (ii) review screens present meaningful text (*i.e.,* candidate names), while receipt review screens have arbitrary strings (*i.e.,* confirmation codes), and (iii) their study measures if manipulations were noticed (based on self-reporting after users are told the review screen was manipulated) while we measure if a voter then files a dispute.

3. RATIONALE AND HYPOTHESES

Assuming that not all individuals who participate in an election will check their votes, checking that a vote has been cast properly may be more likely to occur when there is some level of doubt or uncertainty with the system [Andaleeb 1996; Tolin et al. 2003; van den Hout and Kindt 2003]. That is, individuals might be more likely to check whether their vote was cast properly when some level of uncertainty or discomfort with election results occurs. For example, doubt may be cast if an unexpected or undesired individual is elected. It has been observed previously that social trust is undermined when suspicion is elicited[Lee and Schwarz 2012]. In these cases, we hypothesize individuals who did not vote for the winning party may be more likely to check their receipt.

> H1: Individuals are more likely to check their receipt when the election outcome is unexpected.

A second question addresses how willing voters are to report a discrepancy in the receipt check. Individuals might be more inclined to assume that someone else in the same position would report a problem (the bystander effect), and/or that they, the voter, are at fault, rather than investigate or report that their vote was incorrectly tabulated (a knowledge gap).

> H2: Individuals may be more likely to report a problem to authorities when presented with conflicting/error messages when checking their code.

In summary, we hypothesize that the *relative* vote checking rates and reporting rates will be higher when individuals are presented with unexpected information or errors in tabulation. We are not hypothesizing a specific absolute proportion of voters that may check receipts, only that proportion of voters who do check receipts and report a discrepancy will vary by randomly assigned condition.

4. STUDY 1: EXPECTED VERSUS UNEXPECTED ELECTIONOUTCOMES

The purpose of Study 1 was to test Hypothesis 1. Here, we examine whether individual voters would be more likely to check their receipts when the outcome of the vote was unexpected. We elicited a vote from participants, and then issued them a receipt in form of a ballot ID and confirmation code. Once the outcome of the vote was shared with voters, we examined whether they checked that their receipt was correctly reflected, as well as how perseverant they were in resolving (reporting) a discrepancy. The study was split across two conditions—one where the outcome of the vote was expected, and one where the outcome was unexpected (based on proportion of actual votes obtained).

4.1 Participants

From several case studies on deployments of E2E systems in the literature, there is little consensus on what fraction of voters can be expected to check a receipt—this number has been reported as high as 50% and as low as 4%. We assumed that checking rates would fall somewhere in that range; as such, we anticipated that a sample size of 800 would suffice.

Participants were recruited from Crowdflower (http://crowdflower.com/). Crowdflower is an online crowdsourcing tool, used for recruiting samples for large-scale online surveys. A researcher is able to upload a questionnaire, which is then made available to over 50 labour partner sites. Here, participants are able to access the questionnaire, where they can complete it for a small financial incentive (<$1). The incentive is paid upon completion of the questionnaire; Crowdflower is compensated on a per-participant basis, taking an overhead of payment (33%). Previous work examining similar participant tools, such as Amazon Mechanical Turk, has suggested that the participants involved represent a wide spectrum of individuals with regard to demographics and culture [Paolacci, Chandler, and Ipeirotis 2010]. Other work has demonstrated that these online subject pools provide an inexpensive and reliable source of data [Horton, Rand, and Zeckhauser 2011]. Given that we aim to extrapolate our findings to the general voting public, we chose to use this online pool of participants as our main subject pool.

4.2 Method

In the domain of social psychology it is well understood that simply asking participants to report on their behavior in a novel setting does not always lead to them giving accurate predictions [see Orne, 1962; Seeman, 1969; Weber & Cook, 1972; Weinstein, 1980]. In order to examine actual voting behavior, therefore, we asked participants to take part in a low-stakes voting task. Participants in our study were told that the researchers were interested in charitable donation behavior, and that they would be asked about their own behaviors with regard to charitable giving.

We chose donation behavior as a voting domain because we reasoned that in general, charitable giving is positively valenced. That is, in general, most people feel positively toward charitable giving. Upon reading and signing the consent form (clicking "I Consent"), participants were asked to complete a short questionnaire about their charitable giving behaviors (see Appendix A). Following the questionnaire, participants were given the opportunity to vote for one of five charities and were told that the charity receiving the most votes would be presented with a $50 donation from the experimenters. This voting task represented the main variable of interest in this study. Since a charity vote is relatively low-stakes voting activity, we hypothesize that user trust will be high, and individuals are unlikely to suspect a nefarious outcome. As such, participants may at baseline be unlikely to check their receipt in the same way that voter apathy might impact a political election. In the instructions we decided to reference the potential for "browser compatibility issues" in order to evoke some degree of skepticism in the voting system, and, in turn, to encourage greater incidence of receipt checking. This script was provided for participants in both conditions. The voting task was described in the following text:

> "In addition to receiving payment for participating in our study, we will be donating $50 to a charitable organization. On the next page, you may select one charity from a list; the charity with the most votes will receive the donation. A donation vote is not required; you may simply select "No Thanks" to complete the study.
> After you vote, we will send you an email that will include a confirmation code for your vote, as well the results of the vote. At this point, you can use the confirmation code to check that your vote was correctly included in the result. This is optional, but we recommend it due to compatibility issues that have occurred with certain browsers."

Participants were then taken to the voting page, where they were given the choice to vote for one of the following charities (see Fig. 1).

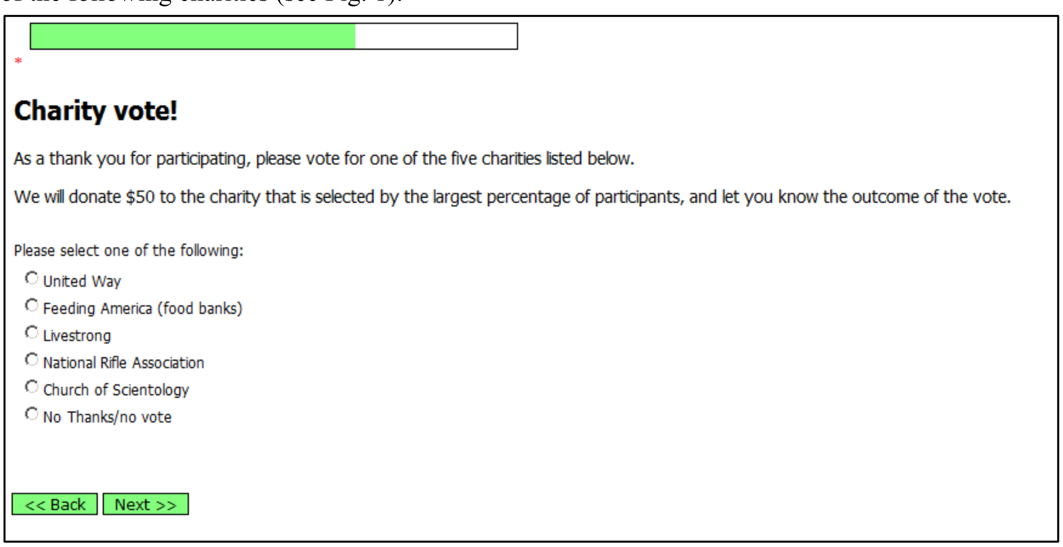
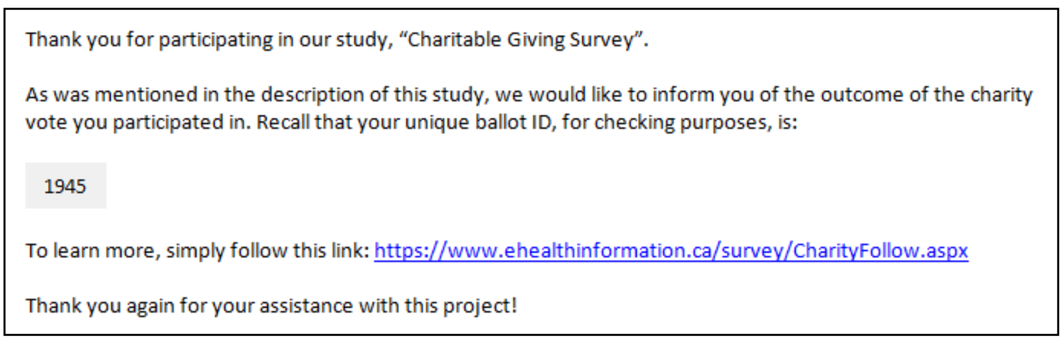


Figure 1. Screenshot of charity vote offered to participants.

We purposely selected some charities that were relatively broad in scope (*i.e.,* United Way, Feeding America/food banks), as well as some that had more narrow interests (*i.e.,* National Rifle Association; Church of Scientology), as well as one associated with recent scandal (*i.e.,* Livestrong's association with Lance Armstrong). Importantly, we did not select "expected" and "unexpected" charity outcomes subjectively; we based these decisions on objective voting behavior of our participants.

After the election closed and all votes were tallied, participants were emailed and thanked for participating. This email included a reminder of each participant's ballot ID and confirmation code, as well as a link to the outcome of the vote (see Fig. 2). For half of the participants (expected outcome condition), the link produced a web page (see Fig. 3) displaying the accurate winner of the vote (i.e., the organization that received the most votes was displayed as the winner). Thus, for this group of participants, the outcome of the vote was highly believable. For the other half of participants (unexpected outcome condition), the web page displayed the winner as the organization that received the least number of real votes. For this group of participants, the outcome of the vote was hypothesized to be less believable, which may have acted as a cue to participants that something suspicious had occurred with the vote. Following this, participants saw the following message:

> "At this point, you may check that your ballot was included in the final result by using the confirmation code we sent previously to you by email, using this link [link to receipt check]." (see Fig. 3).


Thank you for participating in our study, "Charitable Giving Survey".

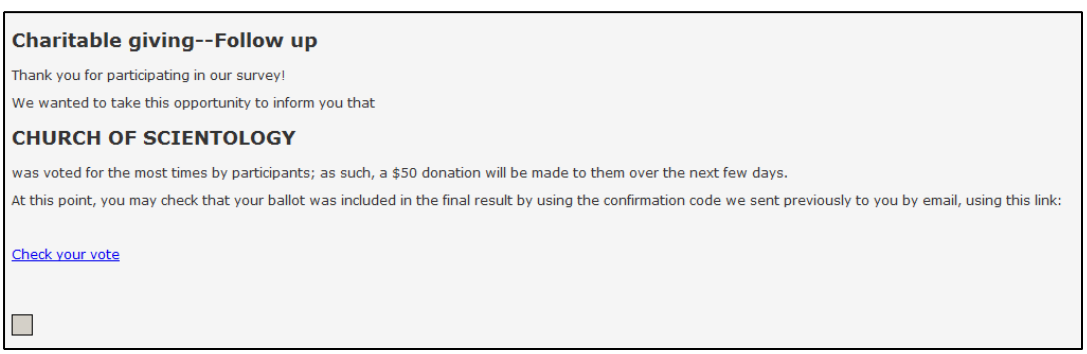
As was mentioned in the description of this study, we would like to inform you of the outcome of the charity vote you participated in. Recall that your unique ballot ID, for checking purposes, is:

1945

To learn more, simply follow this link: https://www.ehealthinformation.ca/survey/CharityFollow.aspx

Thank you again for your assistance with this project!

Figure 2. Screenshot of email sent to participants, informing them of their ballot ID and a link to learn the outcome of the vote.


Charitable giving--Follow up

Thank you for participating in our survey!

We wanted to take this opportunity to inform you that

CHURCH OF SCIENTOLOGY

was voted for the most times by participants; as such, a $50 donation will be made to them over the next few days.

At this point, you may check that your ballot was included in the final result by using the confirmation code we sent previously to you by email, using this link:

Check your vote

Figure 3. Screenshot of the follow-up survey's first page, alerting the participant to the outcome of the vote (in this case, the unexpected outcome).

After clicking the link, participants were prompted to enter their Ballot ID (see Fig. 4). After clicking to "check confirmation code", a confirmation code was then displayed on the following page—this code was either correct or incorrect (i.e., matched or did not match the code that was sent via email). Below the code display, participants saw a link to report an error (see Figs. 5, 6). We hypothesized that individuals would be more likely to check their vote in the unexpected outcome condition (versus the expected outcome condition). Further, we hypothesized that when an incorrect code (versus the correct code) is displayed when checking, participants would be more likely to report an error to authorities.

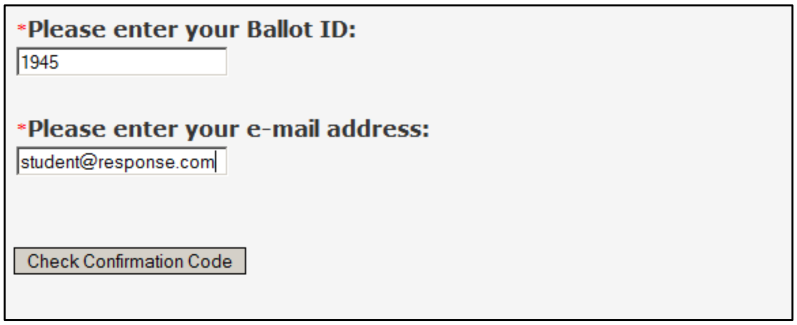


Figure 4. Second page of the follow-up survey, where participants were asked to enter their ballot ID and email address. This form was not pre-filled, but we have included a template for the reader.

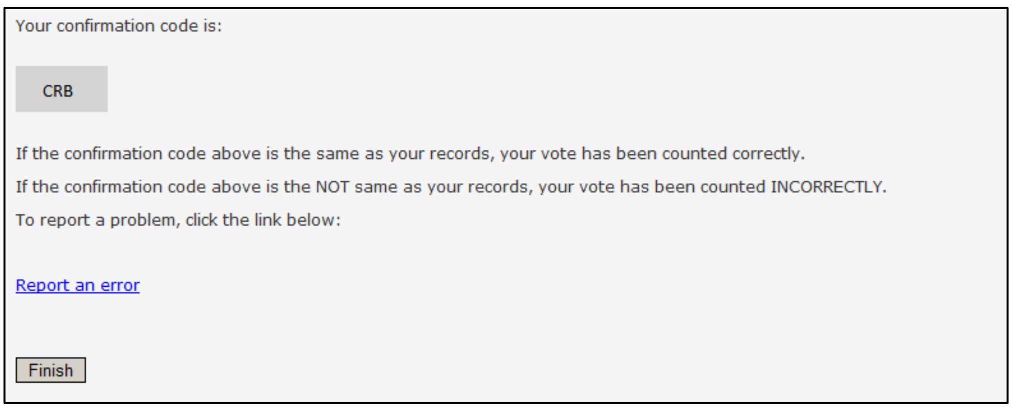


Figure 5. Third page of the follow-up survey, where participants were presented with either a correct or incorrect confirmation code, and given the option to report a problem.

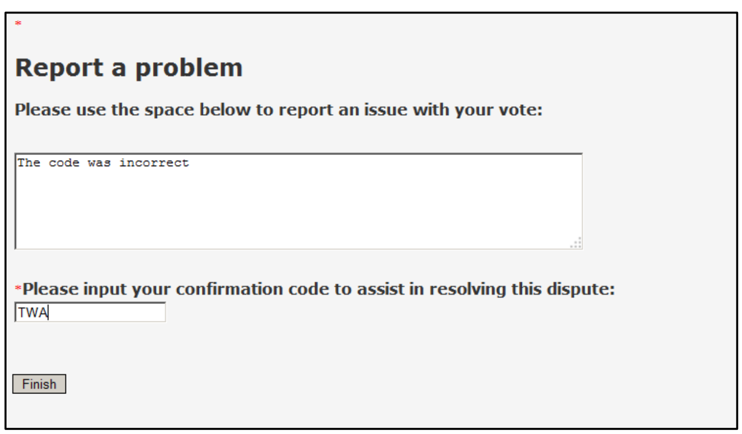


Figure 6. The fifth page of the follow-up survey, where participants were given the option to report a problem with their vote. This form was not pre-filled, but we have included a template for the reader.

Finally, all participants were emailed a copy of the feedback form.

4.3 Results

Participants (N = 841) were recruited from Crowdflower. We collected the following demographic information:

Age. We asked respondents to categorize themselves into one of four age ranges. The majority of respondents fell in the 18-30 year range (48%; non-response: 1.4%). Of the remaining respondents, 32.9% were in the 31-45 age range, 15.3% were in the 46-60 age range, and 1.8% were 60 or older.

Income. We asked respondents to categorize themselves into one of six income brackets The majority of respondents reported an annual income of less than $20,000 (24.7%), or of $20,000 to $39,999 (24.6%; non-response: 7.3%). Of the remaining respondents, 18.1% reported earning $40,000 to $59,000, 11.3% reported earning $60,000 to $79,999, 6.7% reported earning from $80,000 to $99,999, and 7.4% reported earning $100,000 or more. .

Donation behavior. The majority of respondents reported donating once or twice a year (37.6%; non-response: 3.9%). Of the remaining respondents, 5.7% reported never donating, 15.2% reported donating less than once a year, 18.4% reported donating every 2 to 3 months, and 19.3% reported donating at least once a month. Participants tended to donate between $1 and $50 (47.1%; non-response: 4.0%). The majority of participants reported that donating did improve their moods (82.4%; non-response: 4.1%). Of the remaining respondents, 6.1% reported donating nothing, 16.0% reported donating $51 to $100, 10.3% reported donating $101 to $200, 8.2% reported donating $201 to $500, 4.2% reported donating $501 to $1,000, and 4.2% reported donating $1.001 or more.

Lottery winnings. We asked participants whether they would donate lottery winnings to charity, in the event that they won. Most reported that they would donate less than half of their winnings to charity (74.2%; non-response: 7.2%). Of the remaining respondents, 5.7% reported that they would not donate any winnings to charity, 11.7% reported that they would donate more than half of their winnings to charity, and 1.2% reported that they would donate all their winnings to charity. . Participants also reported they were likely to donate this money to multiple charities versus a single charity (76.4%; non-response: 5.6%). Of the remaining respondents, 5.2% said they would not donate, and 12.8% said they would donate to only one charity.

Charity vote. Finally, we asked participants to vote for one of 5 charities for which we would donate $50. The breakdown of votes was as follows:

- United Way: 20.5%
- Feeding America/Food banks: 58.0%
- Livestrong: 6.6%
- National Rifle Association: 6.1%
- Church of Scientology: 1.3%
- No vote: 7.6%

The modal response was for Feeding America/Food banks; a chi-square test confirmed that it was chosen more frequently than all other options, χ^2 (5, 791) = 1072.21, p <.001. As such, we selected

Feeding America/Food banks as our "expected" charity outcome. Similarly, because the Church of Scientology received the fewest number of votes, we selected it as our "unexpected" charity outcome.

Follow-up participation. The last question of the charitable behaviors survey asked participants to leave their email address if they were interested in learning of the results of the vote. Of the 841 total respondents, 603 (71.7%) left their e-mail addresses.

4.4 Vote outcome follow-up

All 603 participants who left their email addresses were contacted with information on the outcome of the charity vote. Participants were emailed a reminder with their unique ballot ID and confirmation code, as well as a link to see the results of the study. The link randomly sent participants to one of two web pages, one of which informed participants of the expected vote outcome (Feeding America/Food banks), and one of which informed participants of the unexpected vote outcome (Church of Scientology). From these pages, participants could then click to check that their vote had been tallied. We measured the proportion of participants who moved on to check whether their vote has been correctly tallied.

Of the 603 participants we emailed, 84 (13.9%) clicked on the link in the email. Of this 84, half of participants received the expected outcome (N = 42; Feeding America/Food banks), and half received the unexpected outcome (N = 42; Church of Scientology). We next examined what proportion of these respondents saw the outcome of the vote and continued to check whether their ballot ID and confirmation codes matched. For each participant, we coded which page they had progressed to[1], and used this value as a continuous variable. A one-way Analysis of Variance revealed that participants in the expected outcome condition (M = 2.52 pages/22 of 42 progressing to ballot ID check) did not progress as far as did those in the unexpected outcome condition (M = 2.76 pages/32 of 42 progressing to ballot ID check), MSE = 1.19, F (1, 83) = 5.39, p = .02. Thus, seeing an unexpected (versus expected) outcome of a vote encouraged more participants to check their votes. These results support our first hypothesis, and suggest that unexpected vote outcomes are likely to produce more receipt-checking behavior, relative to expected vote outcomes. Only three participants (3.6%) clicked the through to see the "report" web page, and only one individual actually filed a report (despite the fact that half of participants who checked their ballot ID and confirmation codes received an incorrect response; N ≈26). This suggests that even when participants are given overt evidence of a discrepancy, few will report a problem. Since we had designed Study 1 only to measure H1, we did not track which condition these participants came from. In response, we formulated H2 and designed Study 2 to specifically test whether individuals presented with unexpected outcomes and incorrect receipts check their receipt in sufficient numbers to detect fraud in an E2E election with high probability.

4.5 Study 1 Discussion

We observed that overall, rates of vote checking were low, though not significantly lower than in previous work[Carback et al. 2010]. Importantly, however, we observed that checking rates were significantly higher when the outcome of the vote was unexpected by voters (H1). Further, though approximately half of participants who checked their ballot IDs and confirmation codes received an incorrect response, only one participant actually filed a report regarding the issue. This suggests

[1] Page 1: Thank you/welcome page; Page 2: Expected vs. unexpected outcome; Page 3: Correct vs. incorrect confirmation code; Page 4: Reporting problem to authorities.

that even when presented with an incorrect code, reporting an error was an unlikely step taken by voters (H2). In Study 2, we specifically test the case where reporting should be highest—that is, when outcomes are unexpected and when the displayed confirmation codes are incorrect.

5. STUDY 2: TARGETED UNEXPECTED OUTCOMES AND INCORRECT CONFIRMATION CODES

In Study 1 we observed that checking rates in E2E voting were greater when the outcome of the vote was unexpected. In the second study we restrict vote outcomes to be unexpected only, in order to maximize the number of samples of voters checking their receipts. Further, we also restrict confirmation codes to be incorrect only. Together, these conditions should encourage vote checking and reporting behaviors.

5.1 Method

As in Study 1, participants were told that the researchers were interested in charitable donation behavior. Upon reading and signing the consent form (clicking "I Consent"), participants were asked to complete a short questionnaire about their charitable giving behaviors. Also, unique to Study 2, participants reported on their online experiences, such as frequency shopping or banking online. Finally, participants were given the opportunity to vote for one of 5 charities for which the experimenters would give a $50 donation. Again, we use the potential for "browser compatibility issues" to evoke some degree of skepticism in the voting system. Participants were then taken to the voting page, where they were given the choice to vote for one of 5 charities (see Fig. 1). After the election closed, and all votes were tallied, participants were emailed and thanked for participating. Included in this email was a reminder of each participant's Ballot ID, as well as a link to learn of the outcome of the vote (see Fig. 2).

All participants were told the winner was the organization that received the least number of actual votes in Study 1 (*i.e.,* Church of Scientology; see Fig. 3). Participants were then given the opportunity to check their receipts as in Study 1 (see Fig. 4). Participants who checked their receipts were told that their confirmation codes were incorrect (See Fig. 5), and were given the opportunity to report the problem (see Fig. 5). Finally, all participants were emailed a copy of the feedback form.

5.2 Results

Participants (N = 755) were recruited from Crowdflower and completed the initial charitable giving survey. We collected the following demographic information:

Time on task. Average time spent on task was 4.12 minutes.

Age. We asked respondents to categorize themselves into one of four age ranges. The majority of respondents fell in the 18-30 year range (50,3%; non-response: 1.9%). Of the remaining respondents, 33.5% were in the 31-45 age range, 13.1% were in the 46-60 age range, and 1.2% were 60 or older.

Income. We asked respondents to categorize themselves into one of six income ranges. The majority of respondents fell across three income brackets, less than $20,000 (20.1%), $20,000 to $39,999 (22.9%), and $40,000 to $59,999 (21.9%; non-response: 9.3%). Of the remaining respondents, 13.9% reported earning $60,000 to $79,999, 6.9% reported earning $80,000 to $99,999, and 5.0% reported earning $100,000 or more.

Donation behavior. The majority of respondents reported donating to charity once or twice a year (34.0%; non-response: 3.7%). Of the remaining respondents, 18.8% reported donating at least once a month, 22.3% reported donating once every 2 to 3 months, 13.9% reported donating less than once a year, and 7.2% reported never donating. Participants reported donating between $1 and $50 (48.5%; non-response: 4.6%). Of the remaining respondents, 7.6% reported that they did not donate, 16.6% reported donating $51 to $100, 10.6% reported donating $101 to $200, 6.6% reported donating $201 to $500, 2.8% reported donating $501 to $1,000, and 2.7% reported donating $1,001 or more. The majority of participants reported that donating did improve their moods (79.5%; non-response: 5.6%).

Lottery winnings. We asked participants whether they would donate lottery winnings to charity, in the event that they won. Most reported that they would donate less than half of their winnings to charity (70.3%; non-response: 7.7%). Of the remaining respondents, 6,5% reported that they would not donate any winnings, 14.1% reported that they would donate more than half of their winnings, and 1.3% reported that they would donate all their winnings. Participants also reported they were likely to donate this money to multiple charities versus a single charity (72.6%; non-response: 5.0%). Of the remaining respondents, 6.5% reported that they would not donate, and 15.8% reported that they would donate to only one charity.

Charity vote. Finally, we asked participants to vote for one of 5 charities for which we would donate $50. The breakdown of votes was as follows:

- United Way: 22.3%
- Feeding America/Food banks: 60.0%
- Livestrong: 6.5%
- National Rifle Association: 4.4%
- Church of Scientology: 1.1%
- No vote: 5.4%

As in Study 1, Feeding America/Food banks was chosen most frequently, and Church of Scientology was chosen least frequently. Again we selected Church of Scientology as our "unexpected" charity outcome.

Follow-up participation. The last question of the charitable behaviors survey asked participants to leave their email address if they were interested in learning the results of the vote. Of the 755 total respondents, 508 (67.3%) left their e-mail addresses.

5.3 Vote outcome follow-up

All 508 participants who left their email addresses were contacted with information on the outcome of the charity vote. Participants were emailed a reminder with their unique ballot ID and confirmation code, as well as a link to see the results of the study. The link sent participants to a web page informing participants of the (unexpected) vote outcome (*i.e.,*Church of Scientology). From this screen participants could click to check their receipt, and we measured the proportion of participants who did so.

Of the 508 participants we emailed, 484 emails were delivered (undelivered emails were due to invalid addresses). Of this sample of 484, 77 clicked on the link in the email. Of this 77, 60 input their ballot ID (57 input a correct ID). Of the 57 participants, only 4 progressed to the Report page. Of the 4 who progressed to the Report page, all 4 filed a report using their confirmation code, and all 4 correctly noted in a comment box that the confirmation code they had previously been given did not match the one provided at check.

We hypothesized that by only providing an unexpected vote outcome we would increase the proportion of individuals who would check their votes. This difference was significant, $t(54) = 3.04$, $p = .004$, suggesting that by targeting participants with an incorrect conformation code, we did indeed increase the percentage of participants who would report a problem (1.8% in Study 1 to 7.3% in Study 2). This report percentage, however, is still quite small—only 0.4% of the voting population checked and reported a problem with their vote.

Lastly, we examined whether propensity to check confirmation codes and report problems was correlated with online experiences (questionnaire in Appendix). Frequency of online shopping and online banking did not correlate with either confirmation code checking ($p > .50$) or reporting a problem ($p > .60$). However, there was a significant correlation observed between number of online accounts they reported as having and propensity to report a problem, $r(754) = .09$, $p = .01$, such that individuals who reported having more online accounts were more likely to report a problem with their vote confirmation codes. This suggests that individuals who interact with more online services could perhaps be more confident in detecting errors, leading to their increased reporting rates.

5.4 Study 2 Discussion

With Study 2, we aimed to encourage vote checking and reporting behaviors. Specifically, we primed participants to be suspicious of a potential error in vote tabulation by informing them that certain browsers had been known to cause trouble for the voting system; we reported the outcome of the charity vote to be the unexpected outcome; and we showed all participants incorrect confirmation codes when they checked their receipt. As such, rates of receipt checking and reporting of errors are likely to be inflated relative to the general population. Compared to Study 1, we observed that more participants filed a report when a discrepancy was present and made explicit (supporting H2). Yet again, however, we observed that rates of receipt checking were low—indeed, likely too low to detect an attack with a high degree of confidence. We observed that 57 of 755 (7.5%) voters checked their receipt and only 4 of 755 (0.5%) reported an incorrect receipt. We posit two possible explanations borrowed from the psychology literature (bystander effect and knowledge gap), which may explain why these receipt checking and reporting rates are lower than one might expect, given previous findings. These results suggest that even extreme cases of potential vote fraud could go undetected by the electorate.

6. GENERAL DISCUSSION

Overall we observed that rates of receipt checking were low, though not significantly lower than in previous work [Carback et al. 2010]. Critically, we observed that checking rates were significantly higher when the outcome of the vote was unanticipated or unexpected by voters in Study 1 (supporting H1). Further, more participants in Study 2 reported an error than in Study 1, suggesting that when incorrect confirmation codes are provided, individuals are more likely to raise an alarm (supporting H2).

Despite this, though approximately half of participants in Study 1, and all participants in Study 2 who checked their ballot IDs and confirmation codes received an incorrect response, only 7 total participants clicked to “report a problem”, and only 5 reported an actual problem, suggesting that even when presented with an incorrect code, reporting an error was an unlikely step taken by voters.

The magnitude of the observed difference between voters who check receipts and those who report errors has significant consequences for E2E elections. Consider a recent close election:

the 2008 United States Senate election in Minnesota, where Sen. Al Franken was certified as the winner by a margin of 225 votes out of 2,887,337. In an election where 7.5% randomly sampled voters confirm their receipts are correct (our observed rate of receipt *checking*), a manipulation capable of altering the result would be detected with 99.99% probability.[2] However, if we reduce the receipt confirmation rate to 0.5% (our observed rate of *error reporting*), the probability of detecting the manipulation is reduced to 43.25%. In other words, it would be more likely to escape detection than not.

Two potential explanations for these inconsistencies in error information and reporting can be drawn from the psychology literature. First, it is possible that voters in this study, though from an online sample and relatively savvy (compared to a population drawn offline), may have had a knowledge gap. Voters may have been confused by the error message, and may have attributed the error to themselves (i.e., thinking they incorrectly remembered or wrote down codes, misunderstanding meaning of error messages, etc.). Alternatively, voters may have experienced a diffusion of responsibility, or the bystander effect [Darley and Latane 1968], such that when faced with a vote tabulation error, they assumed that someone else in the same predicament would report the problem, and so felt as though their reporting was not required. Both of these explanations warrant exploration in future research, as they predict different ways to attenuate the problem: a knowledge gap can be reduced by providing more clear and succinct information; a bystander effect can be reduced by reminding the voter how important their individual vote is, for example.

6.1 Limitations to Study

There are two main limitations to the present studies. First, the task of online voting for a charity donation was lower-stakes relative to the task of voting for a representative in government. As such, rates of reporting errors in the present studies may be lower than in an actual election, and it is possible that receipt checking rates and error reporting would *increase* as the stakes increase.

The second limitation of this work is that participants in this sample were recruited from an online pool and may, therefore, be more technologically adept than the general voting population. As a result, these individuals may be feel more confident in (or less confused by) the receipt checking process, which may have encouraged more checking and error reporting than a typical population. As such, the prevalence of receipt-checking in this work may be *inflated* relative to the general population. Additionally, we made receipt checking relatively easy (by prompting users through email), and error reporting was as simple as filling out a webform (while real elections may require documentation, signed forms, or in-person reporting). These factors may also cause our rates to be inflated.

6.2 Replication Studies

For our observations to be considered reliable when generalized to different elections and voting populations, our hypotheses should be retested in replication studies. We suggest a few extensions and modifications future studies might consider to address the limitations of our own study. To determine to what extent, if any, our results are due to voter apathy concerning the relatively low stakes of our election, follow-up work could implement a similar study with much larger rewards or utilize a poll concerning a real political issue. A follow-up study may also be implemented in a

[2] Assuming an adversary fraudulently changes $F = \lceil 225/2 \rceil$ receipts from Franken to his closest competitor, the probability of detecting fraud with $B = 2887337$ ballots and $\mathrm{E}[R]$ expected reported receipt confirmations is:

$$\mathbf{Pr}_{Detection}[R, B, F] = 1 - \binom{B-F}{\mathrm{E}[R]} \Big/ \binom{B}{\mathrm{E}[R]}$$

real election, however the use of deception makes this course of action questionable both ethically and legally. Future studies should strive to test different voting populations, including ones more representative of the voting-age population.

Future studies may also reexamine all aspects of the study design, and in particular the wording used to inform voters of the option to check their receipt and report an error. We know from other domains in usable security that wording is important (*e.g.,* browser warnings about HTTPS connections [Akhawe and Porter Felt 2013])—different wordings could be tested to see what, if any, measurable impact it has on user behavior. This has been studied to some extent with an emphasis on the communicated message [Olembo et al. 2014].

Finally future studies might examine which voters verify their receipts and report discrepancies, perhaps with a follow-up questionnaire to understand their motivations. The study could also track how these users voted to see if certain voters are more likely to check/report (*e.g.,* supporters of the losing candidates).

6.3 Future Directions for E2E Designers

Future work should examine how voting procedures might be altered to improve usability. In this study, individuals (especially those who are less computer-savvy) may have had difficulty understanding what vote confirmation error messages meant, as well as how to appropriately handle them (i.e., reporting the problem). As such, development of more straightforward communication with regard to receipt checking could increase checking rates, especially among those individuals who would otherwise not check. For example, creating a simple infographic about how to use the system could increase overall vote checking rates. Alternatively, priming individual voters with security concerns could also increase check rates (although likely at the cost of trust). Finally, inconsistent findings in code checking could automatically be reported without requiring the voter to click through to an additional page, or reporting buttons could be made larger and more obvious to voters.

7. CONCLUSIONS

The present work suggests that much more work is required to maximize effectiveness of E2E voting systems, particularly with reference to how people approach vote checking and error reporting. E2E systems may struggle to reliably uncover fraud if the error reporting rates from our studies generalize to real world elections. However, we believe simple design changes could improve usability of E2E voting, thereby improving fraud detection, and thus validity, of the system overall.

8. ACKNOWLEDGEMENTS

We thank the Editorial Board of JETS for useful and actionable feedback that improved the paper. The second author acknowledges NSERC Discovery for partial funding of this work.

9. REFERENCES

Acemyan, C., P.Kortum, M.D. Byrne, and D.S. Wallach. 2014. "Usability of Voter Verifiable, End-to-end Voting Systems: Baseline Data for Helios, Prêt à Voter, and Scantegrity II." *JETS* 2(3).

Adida, B., O. de Marneffe, O. Pereira, and J. J. Quisquater. 2009. "Electing a University President Using Open-Audit Voting: Analysis of Real-World Use of Helios." EVT/WOTE.

Andaleeb, S. S. 1996. "An Experimental Investigation of Satisfaction and Commitment in Marketing Channels: The Role of Trust and Dependence." *Journal of Retailing* 72 (1): 77–93.

Akhawe, D. and A. Porter Felt. 2013. “Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness.” USENIX Security.

Ben-Nun, J., N. Farhi, M. Llewellyn, B. Riva, A. Rosen, A. Ta-Shma, and D. Wikström. 2012. “Wombat: An Implementation and Field Testing of Dual Paper-Cryptographic Election System.” EVOTE.

Carback, R., D. Chaum, J. Clark, J. Conway, A. Essex, P. S. Herrnson, T. Mayberry, et al. 2010. “Scantegrity II Municipal Election at Takoma Park: The First E2E Binding Governmental Election with Ballot Privacy.” USENIX Security.

Clark, J. 2011 “Democracy Enhancing Technologies: Toward deployable and incoercible E2E elections.” PhD Thesis, Waterloo.

Darley, J. M. and B. Latane. 1968. “Bystander Intervention in Emergencies: Diffusion of Responsibility”. *Journal of Personality and Social Psychology* 8 (4, Pt. 1): 377-383.

Essex, A., J. Clark, R. Carback, and S. Popoveniuc. 2007. “Punchscan in Practice: an E2E Election Case Study.” WOTE.

Everett, S. P. 2007. “The Usability of Electronic Voting Machines and How Votes Can Be Changed Without Detection” Dissertation, Rice University.

Horton, J. J., D. G. Rand, and R. J. Zeckhauser. 2011. “The Online Laboratory: Conducting Experiments in a Real Labor Market.” *Experimental Economics* 14 (3): 399–425.

Karayumak, F., M. M. Olembo, M. Kauer, and M. Volkamer. 2011. “Usability Analysis of Helios - An Open Source Verifiable Remote Electronic Voting System.” EVT/WOTE.

Latane, B. and J. M. Darley. 1970. “The Unresponsive Bystander: Why Doesn’t He Help?”.New York, 131 pp.

Lee, S. W. S., and N. Schwarz. 2012. “Bidirectionality, Mediation, and Moderation of Metaphorical Effects: The Embodiment of Social Suspicion and Fishy Smells.” *Journal of Personality and Social Psychology* 103 (5): 737–49.

Olembo M. M., S. Bartsch and M. Volkamer. 2013. “Mental Models of Verifiability in Voting.” VoteID.

Olembo M. M., K. Renaud, S. Bartsch and M. Volkamer. 2014. “Voter, What Message Will Motivate You to Verify Your Vote?” USEC, 2014.

Orne, M. T. 1962. “On The Social Psychology Of The Psychological Experiment: With Particular Reference To Demand Characteristics And Their Implications.” *American Psychologist* 17 (11): 776–83.

Paolacci, G., J. Chandler, and P. G. Ipeirotis. 2010. “Running Experiments on Amazon Mechanical Turk.” *Judgment and Decision Making* 5 (5): 411–19.

Seeman, J. 1969. “Deception In Psychological Research.” *American Psychologist* 24 (11): 1025–28.

Tolin, D. F., J. S. Abramowitz, B. D. Brigidi, and E. B. Foa. 2003. “Intolerance of Uncertainty in Obsessive-Compulsive Disorder.” *Journal of Anxiety Disorders* 17: 233–42.

Van den Hout, M., and M. Kindt. 2003. “Repeated Checking Causes Memory Distrust.” *Behaviour Research and Therapy* 41: 301–16.

Weber, S. J., and T. D. Cook. 1972. “Subject Effects In Laboratory Research: An Examination Of Subject Roles, Demand Characteristics, And Valid Inference.” *Psychological Bulletin* 77 (4): 273–95.

Weinstein, N. D. 1989. “Optimistic Biases about Personal Risks.” *Science* 246 (4935): 1232–34

APPENDIX

A. Charity questionnaire

General demographics

1. Please select an age-range that describes you:
   18-30
   31-45
   46-60
   61-75
   76+
   Prefer not to say

2. In what income bracket do you fall?
   <$20,000
   $20,000 to $39,999
   $40,000 to $59,999
   $60,000 to $79,999
   $80,000 to $99,999
   $100,000+
   Prefer not to say

The following questions ask you about your charitable giving behavior. Please choose the response that best fits your behavior.

1. How frequently do you make a charitable donation (to a registered charity)?
   At least once a month
   Once every 2 or 3 months
   Once or twice a year
   Less than once a year
   Never
   Prefer not to say

2. If you do make a charitable donation, how much do you give on average per year?
   $0—I do not give to charity
   $1—$50
   $51—$100
   $101—$200
   $201—$500
   $501—$1000
   More than $1001
   Prefer not to say

3. If you do make a charitable donation, do you feel better about yourself after donating to charity?
   Yes
   No
   Prefer not to say

4. If you won the lottery, how much of it would you give to charity?

None
Less than half
More than half
All of it
Prefer not to say

5. If you won the lottery and said you would give at least some of it to charity, would you give it all to one charity, or to several?

I said I would not give any of my winnings to charity
I would give to only one charity
I would give to several charities
Prefer not to say

USENIX Journal of Election Technology and Systems (JETS)
Volume 3, Number 1 • December 2014
www.usenix.org/jets/issues/0301

Verifiable European Elections: Risk-limiting Audits for D'Hondt and its relatives

Philip B. Stark
UC Berkeley
stark@stat.berkeley.edu

Vanessa Teague
University of Melbourne
vjteague@unimelb.edu.au

December 5, 2014

Abstract

We provide Risk Limiting Audits for proportional representation election systems such as D'Hondt and Sainte-Laguë. These techniques could be used to produce evidence of correct (electronic) election outcomes in Denmark, Luxembourg, Estonia, Norway, and many other countries.

1 Introduction

Electronic voting in Europe is both controversial and limited. Some countries use polling-place DREs (direct-recording electronic voting machines); others, such as the Netherlands, Ireland, and Germany, introduced and then rejected DREs. No European country requires auditing a paper trail. Some, including Switzerland, Estonia and (until recently) Norway, use Internet voting systems without universally verifiable tallying.

Risk-limiting audits test an announced election result against voter-verified paper records. They aim to answer the question, "Given an agreed list of cast votes, how do we provide convincing public evidence that the election outcome is correct?"[1] The techniques were developed for plurality voting systems. It is not obvious how to adapt them to complex European election systems. This paper fills the gap for many "highest-averages"[2] proportional representation schemes used in Europe, including D'Hondt and Sainte-Laguë. As far as we know, this is the first work to develop risk-limiting audits for highest-averages proportional representation methods.

We provide several RLA techniques for highest-averages elections: If the reported seat allocation is wrong, there is a guaranteed minimum probability that the audit will correct it.

These methods could be used in Norway, Germany, Luxembourg, Estonia, Denmark, Belgium, and other countries. Our work could apply in Belgium, where—after computer scientists pressured the government—the

[1] Ensuring that the list accurately reflects the voters' intentions is also challenging; see, for instance, Stark and Wagner [2012].

[2] This terminology is from Gallagher [1992].

electronic voting machines produce a paper trail, which has never been audited, despite recommendations [BeV, 2007].[3]

1.1 Background and contribution: Risk-limiting audits

We assume we have a voter-verified paper record that has been determined by a *compliance audit* Benaloh, Jones, Lazarus, Lindeman, and Stark [2011], Lindeman and Stark [2012], Stark and Wagner [2012] to reflect the true electoral outcome (The electoral outcome is the number of seats assigned to each party, not the specific number of votes cast for each party.) We also have a *reported* (electronic) outcome, which we distrust.

Risk-limiting audits (RLAs), introduced by Stark [2008a], provide a statistical assurance that the reported outcome matches the *actual* outcome a full hand tally of the paper record would show. If the reported outcome is wrong, no matter why, a risk-limiting audit has a large probability of correcting it. After a RLA, either there is strong statistical evidence that the outcome is correct, or the outcome is known to be correct.

RLAs have been derived for and performed on plurality contests, majority contests, multi-winner contests, and multiple contests simultaneously [Stark, 2008b, Hall, Miratrix, Stark, Briones, Ginnold, Oakley, Peaden, Pellerin, Stanionis, and Webber, 2009, **?**]. Sarwate, Checkoway, and Shacham [2013] consider risk-limiting audits for IRV/STV, Condorcet and Borda. We know of no work on RLAs for highest-averages systems.

We focus on two approaches to RLAs, described by Lindeman and Stark [2012]: *ballot-polling audits*, which rely on the paper ballots but not the electronic record, and *ballot-level comparison audits*, which compare electronic *cast vote records* (tallies for individual ballots) to the corresponding paper records. Both require a *ballot manifest* that describes how ballots are stored. Ballot-polling audits have minimal set-up costs and need nothing from the electronic system except a reported outcome. But they generally involve inspecting more ballots than ballot-level comparison audits, which require that the voting system report results for individual ballots in a way that allows each to be matched to its corresponding paper record—and no federally certified voting system in the US does that. *Batch-level comparison audits*, which compare electronic tallies for bundles of ballots to hand counts of the votes on those ballots, can be performed by substituting the new test statistic we introduce here into existing batch-level RLA methods.

Section 2 develops RLAs for voting schemes in which each voter may cast at most one vote per party, but possibly several votes in all. Section 3 develops a method applicable when voters may cast several votes among different lists. In both sections, we show how to audit which candidates deserve each party's seats, if a simple plurality system is used for that

[3] We do not address auditing the Belgian allocation of party seats to candidates, nor do we develop RLAs for the complex multi-stage seat allocation in Danish and German parliamentary elections—although our methods could form the basis of such audits. As is, the methods can audit the allocation of most of the seats, but not the "compensatory" rounds.

step, as it is in Danish, Luxembourgish, and Norwegian municipal elections. We illustrate the approach using data from the Danish European Parliamentary Election in 2014. Some countries, including Belgium, use a more complicated algorithm for seating candidates within a party—we do not address that audit.

1.2 "Highest-averages" voting methods

"Highest-averages" methods are party-list proportional representation methods: Each voter chooses a party, and the seats are allocated to parties in proportion to the votes each received. Complications arise from rounding, since seats come in integral numbers. (Complications also arise when voters may cast votes for individual candidates or for more than one party. We address such issues in Section 3.) Throughout the paper, we use "party" and "list" interchangeably.

A list of divisors $d(1), d(2), \ldots, d(S)$ determines a highest-averages method. Starting with the tally $t(p)$ for each party $p = 1, \ldots, P$, seats are allocated by calculating $p_{ps} = t(p)/d(s)$ for $p = 1, \ldots, P$ and $s = 1, \ldots, S$. The S seats go to the parties corresponding to the S largest values of p_{ps}, that is, the *winning set* $\mathcal{W}$ is

$$\mathcal{W} = \{(p, s) : t(p)/d(s) \text{ is one of the } S \text{ largest.}\}$$

Every other candidate loses:

$$\mathcal{L} = \{(p, s) \notin \mathcal{W}\}$$

The number of seats assigned to party p is $\#\{(i, s) \in \mathcal{W} : i = p\}$. Some countries compute all $P \times S$ values of p_{ps}, then choose the largest S entries. Others (such as Luxembourg) derive the same result using iterative calculations, known as Jefferson's or Webster's method.

Notation is summarised in Table 1. Table 1.2 shows how seats were allocated to each coalition in the 2014 Danish EU Parliamentary elections, using D'Hondt. Won seats are shown in bold—these were subsequently distributed among coalition members.

If there are more than S values of p_{ps} greater than or equal to the Sth largest, a tie-breaking rule is used to select S of them. In this case the margin is zero and a full hand count is required. Hence we assume from now on that $\#\mathcal{W} = S$ and $\#\mathcal{L} = S(P - 1)$.

"Highest-averages" methods differ in their choice of divisors. Belgium, Denmark, Luxembourg, and many others use the D'Hondt method, for which $d(i) = i$. Sainte-Laguë, which Germany uses, has divisors $1, 3, 5, 7, \ldots$. Estonia and Norway use variants of D'Hondt and Sainte-Laguë respectively.

2 RLAs for one vote per party

Think of each of the $P \times S$ pairs (p, s) as a *pseudo-candidate* reported to have received p_{ps} votes. The set $\mathcal{W}$ contains the reported winners according to the reported tally. The *reported outcome* is the number of

$$
\begin{aligned}
B &: \text{number of ballots cast in the contest} \\
V &: \text{maximum number of votes per ballot} \\
P &: \text{number of parties} \\
S &: \text{number of seats to be assigned} \\
C_p &: \text{number of candidates in party } p \\
t(p) &: \text{reported total for party } p \\
a(p) &: \text{the actual total for party } p \\
e(p) &\equiv t(p) - a(p), \text{ error for party } p \\
t(p,c) &: \text{reported total for candidate } c \text{ in party } p \\
a(p,c) &: \text{actual total for candidate } c \text{ in party } p \\
e(p,c) &\equiv t(p,c) - a(p,c), \text{ error for candidate } c \text{ in party } p \\
d(s) &: \text{the divisor for column } s \\
p_{ps} &\equiv t(p)/d(s) \\
\pi_{ps} &\equiv a(p)/d(s) \\
\mathcal{W} &: \text{the pairs } (p,s) \text{ with the } S \text{ largest values of } p_{ps} \\
\mathcal{L} &: \text{the pairs } (p,s),\ p = 1,\ldots,P,\ s = 1,\ldots,S \text{ not in } \mathcal{W} \\
\mathcal{W}^P &: \text{the parties } p \text{ that (reportedly) won at least one seat} \\
\mathcal{L}^P &: \text{the parties } p \text{ that (reportedly) lost at least one seat} \\
\mathcal{W}_p &: \text{the candidates } c \text{ in party } p \text{ who were seated} \\
\mathcal{L}_p &: \text{the candidates } c \text{ in party } p \text{ who were not seated}
\end{aligned}
$$

Table 1: Notation

	Count in thousands					
Coalition/party	$t(p)$	/2	/3	/4	/5	/6
A+B+F	**833**	**417**	**278**	**208**	**167**	139
Danish People's	**606**	**303**	**202**	**151**	121	101
C+V	**588**	**294**	**196**	147	118	98
People against EU	**184**	92	61	46	37	31
Liberal Alliance	65	33	22	16	13	11

Table 2: Allocating 13 seats among 5 coalitions using D'Hondt, Danish 2014 EU Parliamentary election.

seats each party gets according to the reported totals $t(p)$, $p = 1, \ldots, P$. The *actual outcome* is the number of seats each party would get according to the actual totals $a(p)$, $p = 1, \ldots, P$. The reported outcome is correct if matches the actual outcome, *i.e.*, if and only if

$$\forall (p_w, s_w) \in \mathcal{W},\ \forall (p_\ell, s_\ell) \in \mathcal{L},\ \pi_{p_w s_w} > \pi_{p_\ell s_\ell}, \tag{1}$$

where $\pi_{ps} \equiv a(p)/d(s)$. Auditing consists of checking those $S^2(P-1)$ inequalities statistically. Some of them are entailed by others because $\pi_{ps} > \pi_{pt}$ for $s < t$ for any method with $d(s) < d(t)$. Hence, for instance, if $\pi_{p_w s_w} > \pi_{p_\ell s_\ell}$, then $\pi_{p_w s_w} > \pi_{p_\ell s}$ for all $s \geq s_\ell$, and $\pi_{p_w s} > \pi_{p_\ell s_\ell}$ for all $s \leq s_w$.

For party p, define

$$s_w(p) \equiv \max\{s : (p, s) \in \mathcal{W}\}$$
$$s_\ell(p) \equiv \min\{s : (p, s) \in \mathcal{L}\}.$$

These are the column indices of the last seat p wins and the first seat p loses, respectively. If p won no seats then $s_w(p)$ doesn't exist; if all p's candidates won then $s_\ell(p)$ doesn't exist. At most S parties can have both winners and losers, so at most $\min(2P, S+P)$ of these exist. Define

$$\mathcal{W}^P \equiv \{p : \exists s \text{ s.t. } (p, s) \in \mathcal{W}\}$$
$$\mathcal{L}^P \equiv \{p : \exists s \text{ s.t. } (p, s) \in \mathcal{L}\}.$$

According to the reported results, these are the parties that won at least one seat and the parties that lost at least one seat, respectively. The inequalities that must be checked by auditing are

$$\forall p \in \mathcal{W}^P,\ \forall q \in \mathcal{L}^P \text{ s.t. } p \neq q,\ \pi_{p, s_w(p)} > \pi_{q, s_\ell(q)}. \tag{2}$$

2.1 Ballot-polling Audits

Assumption We assume in this section that the voting rules allow voters to cast at most one vote for at most one party. (A risk-limiting ballot-polling method when voters may cast votes for more than one party or more than one vote per party is given below in Section 3.2.1.)

We will modify the ballot-polling audit method introduced by Lindeman, Stark, and Yates [2012]. Consider a pair of pseudo-candidates $(p_w, s_w) \in \mathcal{W}$ and $(p_\ell, s_\ell) \in \mathcal{L}$, with $p_w \neq p_\ell$. We want to use a random sample to collect and assess evidence regarding whether $\pi_{p_w s_w} > \pi_{p_\ell s_\ell}$. That inequality amounts to $a(p_w)/d(s_w) > a(p_\ell)/d(s_\ell)$, *i.e.*,

$$a(p_w) > a(p_\ell)\frac{d(s_w)}{d(s_\ell)}. \tag{3}$$

Suppose inequality (3) holds. Imagine drawing ballots at random. Let A_p be the event that a randomly selected ballot shows a vote for party p. Then $\Pr(A_p) = a(p)/B$. If the outcome is correct (and if at least one ballot was cast for party p_w or for p_ℓ),

$$\Pr(A_{p_w}) \geq \frac{d(s_w)}{d(s_\ell)} \Pr(A_{p_\ell}),$$

which implies that

$$\Pr(A_{p_w}|A_{p_w} \cup A_{p_\ell}) \geq \frac{d(s_w)}{d(s_\ell)} \Pr(A_{p_\ell}|A_{p_w} \cup A_{p_\ell}),$$

since $A_{p_w} \subset A_{p_w} \cup A_{p_\ell}$ and $A_{p_\ell} \subset A_{p_w} \cup A_{p_\ell}$. That is, the conditional probability $\pi_{p_w|p_w p_\ell}$ that a randomly selected ballot shows a vote for party p_w given that it shows a vote either for p_w or p_ℓ must be at least $d(s_w)/d(s_\ell)$ times the conditional probability $\pi_{p_\ell|p_w p_\ell}$ that such a ballot shows a vote for party p_ℓ. Those two conditional probabilities sum to 100%. Hence, for the outcome to be correct, we need

$$\begin{aligned} \pi_{p_w|p_w p_\ell} &> (1 - \pi_{p_w|p_w p_\ell}) d(s_w)/d(s_\ell) \\ \pi_{p_w|p_w p_\ell}(1 + d(s_w)/d(s_\ell)) &> d(s_w)/d(s_\ell) \\ \text{i.e.,} \quad \pi_{p_w|p_w p_\ell} &> \frac{d(s_w)}{d(s_\ell) + d(s_w)}, \end{aligned} \tag{4}$$

and

$$\pi_{p_\ell|p_w p_\ell} < 1 - \frac{d(s_w)}{d(s_\ell) + d(s_w)}. \tag{5}$$

Now,

$$\pi_{p_w|p_w p_\ell} \equiv \frac{a(p_w)}{a(p_w) + a(p_\ell)}$$

and

$$\frac{t(p_w)}{t(p_w) + t(p_\ell)} > \frac{d(s_w)}{d(s_\ell) + d(s_w)}.$$

We can use Wald's sequential probability ratio test [Wald, 1945] to test the null hypothesis that

$$\frac{a(p_w)}{a(p_w) + a(p_\ell)} \leq \frac{d(s_w)}{d(s_\ell) + d(s_w)}$$

against the alternative hypothesis that

$$\frac{a(p_w)}{a(p_w) + a(p_\ell)} \geq \frac{t(p_w)}{t(p_w) + t(p_\ell)}.$$

To reject the null hypothesis is to confirm that $\pi_{p_w s_w} > \pi_{p_\ell s_\ell}$. In a single draw from the population of ballots, conditional on the event that the ballot shows a vote for either p_w or p_ℓ, the likelihood ratio for the alternative to the null is

$$\frac{\frac{t(p_w)}{t(p_w)+t(p_\ell)}}{\frac{d(s_w(p_w))}{d(s_w(p_w))+d(s_\ell(p_\ell))}}$$

if the ballot shows a vote for p_w. Under the same condition, the likelihood ratio for the alternative to the null is

$$\frac{1 - \frac{t(p_w)}{t(p_w)+t(p_\ell)}}{1 - \frac{d(s_w(p_w))}{d(s_w(p_w))+d(s_\ell(p_\ell))}}$$

if the ballot shows a vote for p_ℓ. Using this likelihood ratio with Wald's sequential probability ratio test [Wald, 1945] gives the following algorithm for an RLA with risk limit α:

1. Select the risk limit $\alpha \in (0,1)$, and M, the maximum number of ballots to audit before proceeding to a full hand count. Define

$$\gamma^+_{ps_w(p)qs_\ell(q)} \equiv \frac{t(p)}{t(p)+t(q)} \cdot \frac{d(s_w(p)) + d(s_\ell(q))}{d(s_w(p))}$$

and

$$\gamma^-_{ps_w(p)qs_\ell(q)} \equiv \quad \left(1 - \frac{t(p)}{t(p)+t(q)}\right) \times \left(1 - \frac{d(s_w(p)) + d(s_\ell(q))}{d(s_w(p))}\right).$$

Set $T_{ps_w(p)qs_\ell(q)} = 1$ for all $p \in \mathcal{W}^P$ and $q \in \mathcal{L}^P$, $p \neq q$. Set $m = 0$.

2. Draw a ballot uniformly at random with replacement from those cast in the contest and increment m.
3. If the ballot shows a valid vote for a reported winner $p \in \mathcal{W}^P$, then for each $q \neq p$ in $\mathcal{L}^P$ that did not receive a valid vote on that ballot multiply $T_{ps_w(p)qs_\ell(q)}$ by $\gamma^+_{ps_w(p)qs_\ell(q)}$. Repeat for all such p.
4. If the ballot shows a valid vote for a reported loser $q \in \mathcal{L}^P$, then for each $p \neq q$ in $\mathcal{W}^P$ that did not receive a valid vote on that ballot, multiply $T_{ps_w(p)qs_\ell(q)}$ by $\gamma^-_{ps_w(p)qs_\ell(q)}$. Repeat for all such q.
5. If any $T_{ps_w(p)qs_\ell(q)} \geq 1/\alpha$, reject the corresponding null hypothesis for each such $T_{ps_w(p)qs_\ell(q)}$. Once a null hypothesis is rejected, do not update its $T_{ps_w(p)qs_\ell(q)}$ after subsequent draws.
6. If all null hypotheses have been rejected, stop the audit: The reported results stand. Otherwise, if $m < M$, return to step 2.
7. Perform a full hand count; the results of the hand count replace the reported results.

Because

$$\frac{t(p_w)}{t(p_w)+t(p_\ell)} > \frac{d(s_w)}{d(s_\ell)+d(s_w)},$$

$Tp_w s_w(p_q)p_\ell s_\ell(p_\ell)$ increases when a ballot with a vote for p_w is drawn and decreases when a ballot for p_ℓ is drawn. If all the alternative hypotheses are true, the values of all the T will tend to increase. If any of the null hypotheses is true, the chance is less than α that the corresponding value of T will ever exceed $1/\alpha$. Hence, as discussed in Lindeman et al. [2012], if any of the null hypotheses is true, despite the fact that we are comparing many pairs of probabilities, there is a large chance that the procedure will require a full hand count: The issue of multiplicity does not arise.

2.2 Comparison audits

Assumption We continue to assume that the voting rules allow voters to cast at most one vote per party, but now we allow votes for multiple parties. (This assumption is relaxed in Section 3.2.2.)

Our approach is similar to the maximum (in-contest) relative overstatement of pairwise margins introduced by Stark [2008b], but with weights in the numerator to account for the fact that a vote for party p amounts to (differing) fractional votes for all the pseudo-candidates in row p.

First, we will transform the problem slightly so that we can use MICRO. We seek a simple sufficient condition for the correctness of the outcome in terms of $e(p)$, $p = 1, \ldots, P$; that is, a condition on the errors in the reported tally that ensures $\pi_{p_w s_w} > \pi_{p_\ell s_\ell}$, $\forall (p_w, s_w) \in \mathcal{W}$ and $\forall (p_\ell, s_\ell) \in \mathcal{L}$.

Suppose there is some $p_\ell \in \mathcal{L}$ and $p_w \in \mathcal{W}$ for which $\pi_{p_w s_w} \leq \pi_{p_\ell s_\ell}$; that is, some seat has been misallocated. Then

$$\begin{aligned} \pi_{p_\ell s_\ell} - \pi_{p_w s_w} &\geq 0 \\ \pi_{p_\ell s_\ell} - p_{p_\ell s_\ell} - (\pi_{p_w s_w} - p_{p_w s_w}) &\geq p_{p_w s_w} - p_{p_\ell s_\ell} \\ \frac{(p_{p_w s_w} - \pi_{p_w s_w}) - (p_{p_\ell s_\ell} - \pi_{p_\ell s_\ell})}{p_{p_w s_w} - p_{p_\ell s_\ell}} &\geq 1. \end{aligned}$$

A little algebra using the definition $p_{ps} \equiv t(p)/d(s)$ shows that the outcome must therefore be correct if

$$\text{MICRO} \equiv \max_{(p_w, s_w) \in \mathcal{W},\, (p_\ell, s_\ell) \in \mathcal{L}} \frac{d(s_\ell)e(p_w) - d(s_w)e(p_\ell)}{d(s_\ell)t(p_w) - d(s_w)t(p_\ell)} < 1.$$

It suffices to take the maximum over $p_w \neq p_\ell$: a party cannot lose a seat to itself.

Let $e_b(p)$ denote the error in the tally of the vote for party p on ballot b. Then $e(p) = \sum_{b=1}^{B} e_b(p)$. Since the sum of maxima dominates the maximum of sums, MICRO < 1 if

$$\sum_{b=1}^{B} \max_{(p_w, s_w) \in \mathcal{W},\, (p_\ell, s_\ell) \in \mathcal{L} : p_w \neq p_\ell} \frac{d(s_\ell)e_b(p_w) - d(s_w)e_b(p_\ell)}{d(s_\ell)t(p_w) - d(s_w)t(p_\ell)} < 1. \tag{6}$$

We now derive a test of hypothesis that MICRO ≥ 1 based on the Kaplan-Wald approach, derived in Appendix A. The test can be modified to use reported results for bundles of ballots rather than individual ballots, at the expense of some bookkeeping; we do not present that generalization here, because for typical bundle sizes and modest margins, it offers little or no advantage over ballot-polling audits, which have far lower set-up costs.

Although one ballot may have been miscounted in a way that affects more than two parties, we need only count the errors that have the largest combined effect on the margin between two pseudo-candidates, because we are summing the maximum effect in the test. Since $|e_b(p)| \leq 1$, the largest possible contribution of any ballot to the left hand side of (6) is

$$u \equiv \max_{w \in \mathcal{W}^P,\, \ell \in \mathcal{L}^P : w \neq \ell} \frac{d(s_\ell(\ell)) + d(s_w(w))}{d(s_\ell(\ell))t(p_w(w)) - d(s_w(w))t(p_\ell(\ell))}. \tag{7}$$

The Kaplan-Wald method requires sampling ballots independently with a probability of selecting each ballot proportional to an upper bound on MICRO for that ballot. Using u as the upper bound on MICRO for *every* ballot results in sampling ballots with equal probabilities—and is conservative.

The following algorithm gives RLA at risk limit α. We assume as before that a compliance audit has shown the audit trail to be sufficiently complete and accurate that a full hand count would show the correct electoral outcome.

The constant γ is a tuning parameter that trades off effort when the cast vote records are error-free against the effort when the cast vote records have errors. The larger γ is (within $[0, 1]$), the smaller the sample will need to be to confirm the outcome when none of the cast vote records is discovered to have error, but the larger the sample will need to be if the audit uncovers errors.

1. Select the risk limit $\alpha \in (0, 1)$; M, the maximum number of ballots to audit before proceeding to a full hand count; and $\gamma \in (0, 1)$. Calculate u and $U = Bu$, the maximum total overstatement. Set $m = 0$.
2. Draw a ballot uniformly at random with replacement from those cast in the contest and increment m.
3. Find MICRO for the selected ballot and divide it by u. Denote the quotient D_m.
4. Calculate $\beta = \prod_{i=1}^{m} \left[\gamma \frac{1-D_i}{1-1/U} + 1 - \gamma\right]$.
5. If $\beta > 1/\alpha$, stop the audit: The outcome is confirmed at risk limit α.
6. If $m < M$, return to step 2.
7. Perform a full hand count; the results of the hand count replace the reported results.

It is a theorem that if any seat was misallocated, the chance this algorithm proceeds to a full hand count is at least $1 - \alpha$: the risk limit is α. Smaller values of γ reduce the increase in workload when discrepancies are found, but increase the workload when no discrepancies are found. The risk limit is conservative regardless.

The method can be simplified and still remain conservative if we replace step 3 by

3') If the selected ballot agrees perfectly with the cast vote record, set $D_m = 0$; otherwise, set $D_m = 1$.

That substitution eliminates the need for any algebra when a discrepancy is discovered, and makes the calculation in step 4 simple. However, it can require inspecting far more ballots when the outcome is correct and discrepancies are observed, because each discrepancy results in multiplying β by $1 - \gamma$.

2.3 Applicability

This method could be used immediately for auditing the number of seats obtained by each list wherever voters may cast only one vote for a list, for example in Danish municipal elections and in Belgium. (See the next section for auditing the candidates assigned to each seat.)

It could also be used in Danish and German parliamentary elections to audit the number of seats obtained by each list in the first (pure) round of D'Hondt tallying. In both countries, the technique would have to be augmented to deal with their complex processes for allocating "compensatory" seats in addition to the D'Hondt count.

Many countries also impose a threshold for parliamentary representation. Some (including Estonia) allow candidates or parties who have exceeded a threshold to be seated immediately, before the D'Hondt count. These could be checked in a straightforward simultaneous audit.

2.4 Illustration: 2014 EU Parliamentary Election in Denmark

Reported results for the 2014 EU Parliamentary election in Denmark are in table 1.2.

An IPython notebook with the data and algorithms is in appendix B and available at XXX. For $\gamma = 0.95$, the allocations of seats to coalitions could have been confirmed at 99.9% confidence ($\alpha = 0.001$ risk limit) by inspecting 1903 ballots—if the audit did not find any errors in that sample.

3 Extension to individual-candidate variants

Many countries allow individual candidate votes. Details vary, but in broad brush, instead of or in addition to choosing a party, voters may select or delete individual candidates. The allocation of seats to parties is as above, based on a combination of party list votes and individual candidate votes. The individual candidate votes are used to decide which candidates in the party are seated.

The electoral outcome can be wrong—the wrong individuals can get seats—either because the parties get the wrong number of seats or because the t candidates within a party that was correctly allocated t seats are not the correct candidates to seat. In many countries, the t candidates in a party who are seated are the t who received the most votes. In that case, we need to test whether every party got the right number of seats and whether, for each party that received at least one seat, the t candidates who reportedly received the most votes really did receive the most votes. The latter amounts to auditing a collection of plurality contests with multiple winners [Stark, 2009]. Below, we extend ballot-polling audits to cover this case.

3.1 Single-list votes plus candidates

In parliamentary elections in Denmark, Belgium, Germany, Estonia, and Norway, voters cast a single party-list vote and may also vote for individual candidate(s) within that list.[4] In these cases, the audit of the seats

[4] This idea is expressed slightly differently in each country. In Denmark, a voter selects either a candidate or a party list. A vote for a candidate is equivalent to a vote for their party list for the purposes of the D'Hondt allocation, but also counts towards that candidate's individual tally for the purposes of assigning seats to candidates within a party. In Germany, voters may select an individual candidate directly (which does not influence the allocation of seats to parties in the Sainte-Laguë count) and then also cast a party-list vote. In Belgium,

allocated per party by the highest-averages count is exactly the same as in the pure case. The audit of which candidates should be seated within each party can be performed simultaneously using the same random sample of ballots by combining tests of pairwise majorities within each party with the test of weighted majorities across parties:

- For ballot-polling, this only requires including additional test statistics for each (seated, non-seated) pair within a party, following Lindeman et al. [2012]: for each pair, we seek strong evidence that the seated candidate received more than half of the votes on ballots that contain votes for either or both candidates.
- For ballot-level comparison audits, we can combine the tests that the seated candidates each received more votes than any of the non-seated candidates by using the maximum across-contest relative overstatement (MACRO) across the pairwise within-party contests, exactly as described by [Stark, 2009]. The Kaplan-Wald method can be used to test the hypothesis that MACRO ≥ 1.

These approaches solve the auditing problem for Denmark and Germany,[5] but not Belgium, which would require a specialized technique tailored to its complicated allocation algorithm.

3.2 Multiple list votes

Assumption This section considers rules that allow a voter to cast multiple votes per party or votes for more than one party. For instance, some countries allow voters to endorse several candidates, who need not be in the same party. For the purposes of a highest-averages method, this is equivalent to giving each voter several votes, which she may distribute among several lists. The highest-averages count then proceeds exactly as in the pure case, except there may be several votes per voter.

The ballot-level comparison RLA of section 2.2 can be modified easily to allow for this possibility—see Section 3.2.2. However, the basic ballot-polling RLA of section 2.1 cannot, even though a comparably simple ballot-polling method works in plurality contests where voters may cast votes for more than one candidate.[6] We develop a different method below in section 3.2.1

Rules vary widely among such systems. For instance, in Luxembourg, voters may choose either a party vote or a candidate vote. In the latter, they may cast up to S votes in total, including up to 2 votes for any single

voters may choose either a party list or an arbitrary number of candidates from the same list, which again is equivalent for the purposes of the D'Hondt tally. Then candidates are seated within parties using a complicated algorithm that combines the voters' and the parties' choices.

[5]This applies only to the first-round of Sainte-Laguë, not the second-round that allocates extra seats in the Bundestag using a different system.

[6]In plurality contests, the basic ballot polling audit checks whether, among ballots that list exactly one of two candidates, one candidate has the majority. Ballots that show both candidates can be ignored. But when a voter can cast more than one vote per party, Party p can have $d(q)/d(p)$ times as many votes as party q among ballots that list exactly one of the two parties, but still not have $d(q)/d(p)$ times as many votes in all, so that conditioning does not yield a valid test.

candidate. Party votes are interpreted as one vote for every candidate on the party list. Party totals are used to allocate seats to parties by D'Hondt; individual votes are used to allocate seats within each party.

Auditing the allocation of seats to candidates requires a method appropriate for the tallying scheme. In Luxembourg this is simple (multi-winner) plurality; in Norwegian municipal elections it is a plurality variant weighted by party selections. Both ballot-polling and ballot-level comparison RLAs can be extended to audit simultaneously how many seats each party gets and which candidates get each party's seats, assuming the latter done by simple plurality. For illustration, we present ballot-polling and comparison audits for the Luxembourgish system.

3.2.1 Ballot-polling audit

Developing a RLA for the Luxembourgish system requires a different approach than that of section 2.1. Voters may cast up to S votes, and up to 2 per candidate, so the probability that a randomly selected ballot shows a vote for a given party or candidate is not proportional to the number of votes for that party or candidate.

This method uses differences in expected values of the number of votes for different parties (normalized by the appropriate column divisors $d(\cdot)$) or for different candidates. We treat a party-list vote as a set of individualised votes for all candidates in that party. Suppose we select a ballot uniformly at random from the B ballots cast. Let $V_{p,c}$ denote the number of votes for candidate c in party p on that ballot and let $V_p \equiv \sum_{c=1}^{C_p} V_{p,c}$ denote the total number of votes for party p on that ballot. Then the expected value of $V_{p,c}$ is

$$\mathbb{E}V_{p,c} = t(p,c)/B \text{ and } \mathbb{E}V_p = t(p)/B.$$

Moreover,

$$\mathbb{E}(V_p/d(s) - V_q/d(t)) = \frac{t(p)/d(s) - t(q)/d(t)}{B}. \tag{8}$$

The allocation of seats to parties is therefore correct if

$$\forall p \in \mathcal{W}^P,\ \forall q \in \mathcal{L}^P \text{ s.t. } p \neq q,\ \mathbb{E}\left(\frac{V_p}{d(s_w(p))} - \frac{V_q}{d(s_\ell(q))}\right) > 0. \tag{9}$$

The allocation of seats to candidates in those parties is also correct if

$$\forall p \in \mathcal{W}^P,\ c_w \in \mathcal{W}_p,\ c_\ell \in \mathcal{L}_p,\ \mathbb{E}(V_{p,c_w} - V_{p,c_\ell}) > 0. \tag{10}$$

Voting rules for a particular country impose constraints that imply lower and upper bounds on the combinations of random variables on the left-hand sides of (9) and (10). Let X_i denote any of those left-hand sides, calculated for the ith draw. (Draws are random, independent, and uniformly distributed.) Let x_+ and x_- denote the upper and lower bounds respectively. For example, in the Luxembourgish system, the rules require $V_{p,c} \leq 2$ and $V_p \leq S$, so if X_i denotes

$$\frac{V_p}{d(s_w(p))} - \frac{V_q}{d(s_\ell(q))}$$

for the ith draw (Eq. (9)), then x_+ is $S/d(s_w(p))$ and x_- is $-S/d(s_\ell(q))$.

We know a priori that $x_- \leq X_i \leq x_+$; we wish to test the hypothesis that $\mathbb{E}X_i \leq 0$. Rejecting that hypothesis for all the left-hand sides confirms the seat allocation. Let $\tilde{X}_i \equiv (X_i + x_-)/(x_+ - x_-)$. Then $\tilde{X}_i \in [0, 1]$, and the condition $\mathbb{E}X_i \leq 0$ is equivalent to the condition $\mathbb{E}\tilde{X}_i \leq t \equiv x_-/(x_+ - x_-)$. Imagine drawing n ballots, resulting in $\{\tilde{X}_i\}_{i=1}^n$ independent and identically distributed on $[0, 1]$. Define

$$\mathrm{LR} \equiv \prod_{i=1}^{n} \left[\gamma \frac{\tilde{X}_i}{t} + 1 - \gamma \right]. \tag{11}$$

Much the same proof as in appendix A[7] shows that if $\mathbb{E}\tilde{X}_i \leq t$, $\Pr\{\mathrm{LR} > 1/\alpha\} \leq \alpha$ for any n. We can use this result to test all the conditions (9) and (10) with a single sample. Multiplicity is not a concern because the audit proceeds to a full hand count if *any* null hypothesis is not rejected.

3.2.2 Ballot-level comparison audit

The algorithm of section 2.2 can audit the number of seats allocated to parties in the case of allowing up to V votes per party per voter, except that the upper bound u_m on the maximum possible value of MICRO for a single ballot is V times as large as that in Equation 7. For Luxembourg, the maximum votes per ballot is the number of available seats, so $u_m = Su$. To include the competition for seats among members of the same party, we need only consider that competition to be a collection of pairwise elections between all candidates in a party who were awarded seats and all who were not. Table 1 outlines the notation.

The definition of MACRO incorporating both kinds of error is:

$$\mathrm{MACRO}_{\mathrm{multi}} \equiv \max \left\{ \mathrm{MICRO}, \max_{p \in \mathcal{W}^P, c_w \in \mathcal{W}_p, c_\ell \in \mathcal{L}_p} \frac{e(p, c_w) - e(p, c_\ell)}{t(p, c_w) - t(p, c_\ell)} \right\}.$$

If $\mathrm{MACRO}_{\mathrm{multi}} < 1$, the allocation of seats to parties and the allocation of seats to candidates within parties are all correct.

3.2.3 Logistical and statistical concerns

If relatively few votes separate a seated candidate from a candidate in the same party who is not seated, the sample sizes needed to attain reasonable risk limits using the methods presented above will be very large. If it is possible to divide the ballots into (overlapping) subsets that contain only the ballots cast for a particular party, and to sample directly from those subsets, it may be possible to reduce sample sizes, depending on the margins compared to the number of ballots in each subset. Auditing the allocation of seats within parties separately from auditing the allocation of seats to parties also raises issues of multiple testing, which will tend to increase the required sample size to attain a given risk limit.

[7] The proof appears in sketch form in `http://printmacroj.com/martMean.htm`, last accessed 10 November 2013.

3.3 Audit Summary

We have presented ballot-polling and ballot-level comparison RLAs for all highest-averages proportional representation methods, including those in which voters select a single party list and those in which they may cast some votes for each of several parties and more than one vote for the same party or candidate. We have shown that the same sample can be used to check that the right candidates were seated within each party, at least for (the many) countries that use plurality or a simple variant to allocate seats to candidates. The methods need modifications to check the "compensatory" rounds in German and Danish parliamentary elections, which do not use a highest-averages method, and for auditing which candidates get the party's seats in non-plurality systems such as Belgian and Norwegian parliamentary elections.

4 Conclusion

Highest-averages methods include many party-list proportional representation methods, implemented differently in different countries—and sometimes in different ways in a single country. The pure versions of these methods are amenable both to efficient risk-limiting audits and to complete homomorphic tallying. We develop methods for several variants, some of which are particularly important because the country uses or plans to use electronic voting. In particular, we illustrate risk-limiting audits for Denmark and privacy-preserving universally verifiable tallying for Norway. The methods allow election outcomes of D'Hondt, Sainte-Laguë, and variants to be verified.

5 Acknowledgements

Many thanks to Christian Bull, Sven Heiberg, Carsten Schürmann, Kristian Gjøsteen, Olivier Pereira, Melanie Volkamer, Peter Ryan and Rui Joaquim.

References

Bevoting study of electronic voting systems, 2007. Available at `http://www.ibz.rrn.fgov.be/fileadmin/user_upload/Elections/fr/presentation/bevoting-2_gb.pdf`.

J. Benaloh, D. Jones, E. Lazarus, M. Lindeman, and P.B. Stark. Soba: Secrecy-preserving observable ballot-level audit. In *proc. Proc. USENIX Accurate Electronic Voting Technology Workshop*, 2011.

M. Gallagher. Comparing proportional representation electoral systems: Quotas, thresholds, paradoxes and majorities. *British Journal of Political Science*, 22(4):469–496, 1992.

J.L. Hall, L.W. Miratrix, P.B. Stark, M. Briones, E. Ginnold, F. Oakley, M. Peaden, G. Pellerin, T. Stanionis, and T. Webber. Implementing risk-limiting post-election audits in California. In *Proc. 2009 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE '09)*, Montreal, Canada, August 2009. USENIX. URL `http://www.usenix.org/event/evtwote09/tech/full_papers/hall.pdf`.

H.M. Kaplan. A method of one-sided nonparametric inference for the mean of a nonnegative population. *The American Statistician*, 41:157–158, 1987.

M. Lindeman and P.B. Stark. A gentle introduction to risk-limiting audits. *IEEE Security and Privacy*, 10:42–49, 2012.

M. Lindeman, P.B. Stark, and V. Yates. BRAVO: Ballot-polling risk-limiting audits to verify outcomes. In *Proceedings of the 2011 Electronic Voting Technology Workshop / Workshop on Trustworthy Elections (EVT/WOTE '11)*. USENIX, 2012.

Panel on Nonstandard Mixtures of Distributions. *Statistical models and analysis in auditing: A study of statistical models and methods for analyzing nonstandard mixtures of distributions in auditing.* National Academy Press, Washington, D.C., 1988.

A.D. Sarwate, S. Checkoway, and H. Shacham. Risk-limiting audits and the margin of victory in nonplurality elections. *Politics, and Policy*, 3 (3):29–64, 2013.

P.B. Stark. Conservative statistical post-election audits. *Annals of Applied Statistics*, 2008a.

P.B. Stark. A sharper discrepancy measure for post-election audits. *Ann. Appl. Stat*, 2:982–985, 2008b.

P.B. Stark. Efficient post-election audits of multiple contests: 2009 California tests. `http://ssrn.com/abstract=1443314`, 2009. 2009 Conference on Empirical Legal Studies.

P.B. Stark and D.A. Wagner. Evidence-based elections. *IEEE Security and Privacy*, 10:33–41, 2012.

A. Wald. Sequential tests of statistical hypotheses. *Ann. Math. Stat.*, 16: 117–186, 1945.

A. Wald. *Sequential Analysis.* Dover Publications, Mineola, New York, 2004.

A The Kaplan-Wald Method

This section combines ideas from dollar-unit sampling as used in financial auditing [Panel on Nonstandard Mixtures of Distributions, 1988] with a technique described in H.M. Kaplan's website, `http://printmacroj.com/martMean.htm`.[8] Kaplan's work fleshes out an idea due to Wald [Wald, 1945, 2004], and is closely related to a technique presented in Kaplan [1987]. We have a population of N items. Item j has a value x_j between 0 and a known upper bound $u_j > 0$. We wish to estimate the population total $T = \sum_{j=1}^{N} x_j$.

Define $d_j \equiv x_j/u_j$, for $j = 1, \ldots, N$. Each d_j is necessarily between 0 and 1. Let $U = \sum_{j=1}^{N} u_j$. We will make n independent random draws with replacement from the population; the probability of selecting item j is $p_j \equiv u_j/U$ in each draw. That is, the chance of selecting item j is proportional to its upper bound.

Let $J(i)$ be the index of the item selected on the ith draw. Let D_i be $d_{J(i)}$, the value of d for the item selected on the ith draw. For instance, if the second draw gives the fifth item, then $J(2) = 5$ and $D_2 = d_{J(2)} = d_5$. The chance that $J(i) = j$ is p_j. The expected value of D_i is

$$
\begin{aligned}
\mathbb{E} D_i &= \sum_{j=1}^{N} d_j \Pr(J(i) = j) \\
&= \sum_{j=1}^{N} d_j (u_j/U) \\
&= \sum_{j=1}^{N} (x_j/u_j) \times (u_j/U) \\
&= \sum_{j=1}^{N} x_j/U \\
&= T/U.
\end{aligned}
$$

Hence,

$$
\mathbb{E} \frac{1}{n} \sum_{i=1}^{n} D_i = \frac{1}{n} nT/U = T/U. \tag{12}
$$

That is, the average of the n draws $\{D_i\}_{i=1}^{n}$ is an unbiased estimator of the population total T as a fraction of the total upper bound U. Equivalently, U times the average of $\{D_i\}_{i=1}^{n}$ is an unbiased estimate of the population total T.

The expected value of D_i is T/U, which is unknown since T is unknown. For the purpose of conducting a risk-limiting audit, we want to test the hypothesis that $T \geq 1$ (equivalently, that $T/U \geq 1/U$). We will derive a method based on Wald's sequential probability ratio test [Wald, 1945, 2004], following an idea of Harold Kaplan, based in turn on a remark in Wald [2004].[9] Note that if $U < 1$, we do not need to audit: the

[8] Last accessed 10 November 2013.

[9] See `http://printmacroj.com/martMean.htm`. Last accessed 10 November 2013.

maximum possible value of MICRO is less than 1, so the outcome must be correct. We therefore assume that $U \geq 1$.

The likelihood ratio of the simple hypothesis H_1 to the simple hypothesis H_0 is the probability of observing the data that actually were observed on the assumption that H_1 is true, divided by the probability of observing the data that were actually observed on the assumption that H_0 is true:

$$\text{likelihood ratio} \equiv \frac{\Pr(\text{observed data if } H_1 \text{ is true})}{\Pr(\text{observed data if } H_0 \text{ is true})}. \tag{13}$$

The probability of observing the data actually observed will tend to be higher for whichever hypothesis is in fact true, so the likelihood ratio will tend to be greater than 1 if H_1 is true, and will tend to be less than 1 if H_0 is true. The more observations we make, the more probable it is that the resulting likelihood ratio will be small if H_0 is true. Wald [1945] showed that if H_0 is true, then the probability is at most α that the likelihood ratio is ever greater than $1/\alpha$, no matter how many observations are made.

Let $\tilde{D}_i \equiv 1 - D_i$, and let $\tilde{d}_j \equiv 1 - d_j$. Then the probability distribution of $\tilde{D}_i$ is

$$f(d) \equiv \sum_{j=1}^{N} p_j \delta(d - \tilde{d}_j)),$$

where $\delta(\cdot)$ is the Dirac delta function. Under the hypothesis that $T = t$, the expected value of $\tilde{D}_i$ is $1 - t/U$, so the expected value of $(1 - t/U)^{-1}\tilde{D}_i$ is 1. That is,

$$\int_{d=0}^{1} (1 - t/U)^{-1} d f(d) = \sum_{j=1}^{N} (1 - t/U)^{-1} \tilde{d}_j p_j = 1. \tag{14}$$

Let $\gamma \in [0, 1]$ be a fixed number. Because $\sum_j p_j = 1$, it follows that if $T = t$,

$$\begin{aligned}
&\mathbb{E}(\gamma(1 - t/U)^{-1}\tilde{D}_i + (1 - \gamma)) \\
&= \frac{\gamma}{1 - t/U} \mathbb{E}\tilde{D}_i + (1 - \gamma) \\
&= \frac{\gamma}{1 - t/U} (1 - t/U) + 1 - \gamma \\
&= \gamma \cdot 1 + (1 - \gamma) \\
&= 1.
\end{aligned}$$

Now,

$$\mathbb{E}(\gamma(1 - t/U)^{-1}\tilde{D}_i + (1 - \gamma)) \equiv \sum_{j=1}^{N} (\gamma(1 - t/U)^{-1}\tilde{d}_j + 1 - \gamma) p_j. \tag{15}$$

Let

$$g_{j,t,\gamma} \equiv (\gamma(1 - t/U)^{-1}\tilde{d}_j + 1 - \gamma) p_j, \quad j = 1, \ldots, N. \tag{16}$$

Since $t/U \in [0,1]$ and all $\{\tilde{d}_j\}$ are nonnegative, it follows from (15) and (**??**) that $g_{j,t,\gamma} \geq 0$ and

$$\sum_{j=1}^{N} g_{j,t,\gamma} = 1. \tag{17}$$

That is, $\sum_{j=1}^{N} g_{j,t,\gamma}\delta(d - \tilde{d}_j)$ is a probability distribution. Let F be a random variable with $\Pr\{F = \tilde{d}_j\} = g_{j,t,\gamma}$. Since $\mathbb{E}\tilde{D}_i = 1 - t/U \geq 0$,

$$\begin{aligned}
\mathbb{E}F &= \sum_{j=1}^{N} \Big(\gamma(1 - t/U)^{-1}\tilde{d}_j + 1 - \gamma\Big)\, \tilde{d}_j p_j \\
&= \frac{\gamma}{1 - t/U} \sum_{j=1}^{N} \tilde{d}_j^2 p_j + (1-\gamma)\mathbb{E}\tilde{D}_i \\
&= \frac{\gamma}{1 - t/U} \mathbb{E}\tilde{D}_i^2 + (1-\gamma)\mathbb{E}\tilde{D}_i \\
&\geq \frac{\gamma}{1 - t/U} (\mathbb{E}\tilde{D}_j)^2 + (1-\gamma)\mathbb{E}\tilde{D}_i \\
&= \gamma\mathbb{E}\tilde{D}_i + (1-\gamma)\mathbb{E}\tilde{D}_i = \mathbb{E}\tilde{D}_i,
\end{aligned}$$

where the penultimate step follows from Jensen's inequality.

If the data allow us to reject the hypothesis H_0 that $\{\tilde{D}_i\}$ all have the same probability mass function f (for which $\mathbb{E}\tilde{D}_i = 1 - t/U$) in favor of the alternative hypothesis H_1 that $\{\tilde{D}_i\}$ all have the probability mass function $g_{t,\gamma}$ (for which $\mathbb{E}\tilde{D}_i > 1 - t/U$), we have strong statistical evidence that $\mathbb{E}D_i < t/U$. Since $\mathbb{E}D_i < t/U$ is a sufficient condition for the electoral outcome to be correct, rejecting H_0 means the audit can stop: The data gave strong evidence that the election outcome is correct.

Recall that $J(i)$ is the index of the item selected on the ith draw. For n independent observations $\{\tilde{D}_i\}_{i=1}^{n}$, the likelihood ratio of H_1 to H_0 is

$$\begin{aligned}
\mathrm{LR} &= \frac{\Pr(\text{observed data if } H_1 \text{ is true})}{\Pr(\text{observed data if } H_0 \text{ is true})} \\
&= \frac{\prod_{i=1}^{n} \Big[\gamma(1 - t/U)^{-1}\tilde{D}_i + 1 - \gamma\Big]\, p_{J(i)}}{\prod_{i=1}^{n} p_{J(i)}} \\
&= \prod_{i=1}^{n} \left[\gamma \frac{1 - D_i}{1 - t/U} + 1 - \gamma\right].
\end{aligned} \tag{18}$$

The dependence on $\{p_j\}$ in the numerator and denominator cancel fortuitously: The validity of the test does not depend on any assumptions about the population $\{d_j\}$ of values. Equation (18) motivates the introduction of γ: For $\gamma = 1$, the likelihood ratio would forever be 0 if even a single observed value of D_i were equal to 1.

To conduct a risk-limiting audit, we take $t = 1$ in (18). If in fact $T \geq 1$, Wald's sequential probability ratio test establishes that the chance that the likelihood ratio is ever larger than $1/\alpha$ is at most α, no matter what the population $\{d_j\}$ of values may be. If we continue to inspect ballots until $\mathrm{LR} > 1/\alpha$—or until we have inspected all the ballots—the chance the audit will stop short of a full hand count if the outcome is wrong is less than α.

B Python code for the European Union Parliamentary election in Denmark

```
import math
import numpy as np
import scipy
from scipy.stats import binom
import pandas as pd
#
def dHondt(partyTotals, seats, divisors):
    '''
    allocate <seats> seats to parties according to <partyTotals> votes,
    using D'Hondt proportional allocation with <weights> divisors

    Input:
        partyTotals: list of total votes by party
        seats:       total number of seats to allocate
        divisors:    divisors for proportional allocation.
                     For d'Hondt, divisors are 1, 2, 3, ...

    Returns:
        partySeats:  list of number of seats for each party
        seated:      list of tuples--parties with at least one seat,
                     number of votes that party got,
                     and divisor for last seated in the party
        notSeated:   list of tuples--parties with at least one lost seat,
                     number of votes that party got,
                     and divisor for the first non-seated in the party
        pseudoCandidates: matrix of votes for each pseudocandidate
    '''
    pseudoCandidates = np.array([partyTotals,]*seats, ).T/divisors.astype(float)
    sortedPC = np.sort(np.ravel(pseudoCandidates))
    lastSeated = sortedPC[-seats]
    theSeated = np.where(pseudoCandidates >= lastSeated)
    partySeats = np.bincount(theSeated[0], minlength=len(partyTotals))
                                       # number of seats for each party
    inx = np.nonzero(partySeats)[0]    # only those with at least one seat
    seated = zip(inx, partyTotals[inx], divisors[partySeats[inx]-1])
                                       # parties with at least one seat,
                                       # number of votes that party got,
                                       # and divisor for last seated in
                                       # the party
    theNotSeated = np.where(pseudoCandidates < lastSeated)
    partyNotSeats = np.bincount(theNotSeated[0], minlength=len(partyTotals))
                                       # number of non-seats for each
                                       # party
    inx = np.nonzero(partyNotSeats)[0]
    notSeated = zip(inx, partyTotals[inx], divisors[partySeats[inx]])
```

```
                                            # parties with at least one
                                            # unseated, number of votes
                                            # for the first non-seated
                                            # in the party
    if (lastSeated == sortedPC[-(seats+1)]):
        raise ValueError("Tied contest for the last seat!")
    else:
        return partySeats, seated, notSeated, lastSeated, pseudoCandidates

def uMax(win, lose):
    '''
    finds the upper bound u on the MICRO for the contest
    win and lose are lists of triples: [party, tally(party), divisor]
    the divisor for win is the largest divisor for any seat the party won
    the divisor for lose is the smallest divisor for any seat the party lost
    See Stark and Teague, 2014, equations 4 and 5.

    Input:
        win:  list of triples--party, tally(party), divisor
        lose: list of triples--party, tally(party), divisor

    Returns:
        maximum possible relative overstatement for any ballot
    '''
    u = 0.0
    for w in win:
        for ell in lose:
            if w[0] != ell[0]:
                u = max([u,
                  (float(ell[2]) + float(w[2]))/float(ell[2]*w[1] - w[2]*ell[1])])
    return u

def minSampleSize(ballots, u, gamma=0.95, alpha=0.1):
    '''
    find smallest sample size for risk-limit alpha, using cushion gamma \in (0,1)
    1/alpha = (gamma/(1-1/(ballots*u))+1-gamma)**n
    Input:
        ballots: number of ballots cast in the contest
        u:       upper bound on overstatement per ballot
        gamma:   hedge against finding a ballot that attains the upper bound.
                 Larger values give less protection
        alpha:   risk limit
    '''
    return math.ceil(math.log(1.0/alpha) /
                     math.log(gamma/(1.0-1.0/(ballots*u)) + 1.0 - gamma))

# final 2014 Danish EU Parliamentary election results from
#   http://www.dst.dk/valg/Valg1475795/valgopg/valgopgHL.htm
# there were two coalitions: (A,B,F) and (C,V)
# There were 13 seats to allocate.
```

```
#
#  Official results by party
#
A = 435245
B = 148949
C = 208262
F = 249305
I = 65480
N = 183724
O = 605889
V = 379840
Ballots = 2332217  # includes invalid and blank ballots
nSeats = 13  # seats to allocate
#
# allocate seats to coalitions
#
coalitionTotals = np.array([A+B+F, C+V, I, N, O])  # for coalitions
coalitionSeats, coalitionSeated, coalitionNotSeated, coalitionLastSeated,
                coalitionPCs
                = dHondt(coalitionTotals, nSeats, np.arange(1, nSeats+1))
print 'A+B+F, C+V, I, N, O:', coalitionSeats
#
# allocate seats within coalitions
#
nABFSeats = coalitionSeats[0]
nCVSeats = coalitionSeats[1]
ABFSeats, ABFSeated, ABFNotSeated, ABFLastSeated, ABFPCs
         = dHondt(np.array([A, B, F]), nABFSeats, np.arange(1, nABFSeats+1))
CVSeats, CVSeated, CVNotSeated, CVLastSeated, CVPCs
         = dHondt(np.array([C, V]), nCVSeats, np.arange(1, nCVSeats+1))
#
print 'A, B, F:', ABFSeats, ';  C, V:', CVSeats
#
ASeats = ABFSeats[0]
BSeats = ABFSeats[1]
CSeats = CVSeats[0]
FSeats = ABFSeats[2]
ISeats = coalitionSeats[2]
NSeats = coalitionSeats[3]
OSeats = coalitionSeats[4]
VSeats = CVSeats[1]
allSeats = [ASeats, BSeats, CSeats, FSeats, ISeats, NSeats, OSeats, VSeats]
print '---------------\nSeats to parties A, B, C, F, I, N, O, V: ', allSeats
print 'Seated coalitions, votes, divisor:', coalitionSeated
print 'Non-Seated coalitions, votes, divisor:', coalitionNotSeated
#
#  Set audit parameters
gamma = 0.95  # tuning constant in the Kaplan-Wald method
alpha = 0.001 # risk limit
#
```

```
u = uMax(coalitionSeated, coalitionNotSeated)
print Ballots*u
n = math.ceil(math.log(1.0/alpha) /
              math.log(gamma/(1.0-1.0/(Ballots*u)) + 1.0 - gamma))
print n
```