## 2006 USENIX/ACCURATE Electronic Voting Technology Workshop (EVT '06)

*Vancouver, B.C., Canada*
*August 1, 2006*

### KEYNOTE ADDRESS

- ***Who Won? Statistical Election Fraud Detection and Its Limits***

  *Walter R. Mebane, Jr., Cornell University*

  *Summarized by Sarah P. Everett*

  Professor Walter Mebane began his work in quantitative political analysis in 2000, after the election fiasco in Florida. He developed methods to detect statistical anomalies in voting like those that occurred owing to the use of the butterfly ballot. After the problems in the 2004 election in Ohio, Mebane expanded his work to use statistical anomalies to detect fraud in elections. He looked at patterns in various states, including Ohio and Florida, where he examined the problems caused by the new electronic voting machines.

  Mebane's project required him to develop new statistical methods. To use the previous methods, much information would have been needed about the election. However, using a new tool based on Benford's Law, all one needs to examine anomalies are the actual votes. Mebane has applied this method to the recent Mexico election data.

  Many political scientists became interested in the usability of voting ballots and equipment after the Florida butterfly-ballot recount. To examine how the butterfly ballot affected that election, Mebane fit a statistical model to votes in many counties. For example, the method allows a researcher to look at the probability that someone voted for Buchanan. This model takes into account voter registration by political party and demographics to predict how many votes Buchanan should have gotten in the county. When the model does not approximate the reported votes, it suggests that something abnormal happened in the county. For the presidential race, box plots of studentized residuals can be made and outliers identified. In Mebane's analysis of the 2000 election data, Palm Beach County was approximately 35 standard deviations away from the rest of the data. This means that the county's count was not produced in the same way as those of the rest of the counties in Florida. In another comparison, absentee ballots (which were not butterfly ballots) were compared to election day ballots to see the percentages who voted for Buchanan. Again, Palm Beach County pops out of the data.

  In Mebane's publication "The Wrong Man Is President," he looks at votes and overvotes in the 2000 Florida election. Some overvotes reflect confusion on the part of the voter. This confusion can be caused by, for example, ballots that instruct voters to vote on every page, although a race may be split between pages. In this election, two-mark and multiple-mark overvotes were enough to make up the difference between the totals for Bush and Gore. In Florida, many types of punchcards and optical scan ballots were used. One can look at the ratio of allocated ballots to certified vote counts to see the rate of overvoting in different counties and on different technologies. Mebane explored how many of the two-mark and multiple-mark overvotes were errors. He used a method that looked at Senate votes and produced true votes. Without the overvote errors, he found, Gore would have gained over 46,000 votes and Bush would have gained approximately 11,000 votes. This led to his belief that the wrong man became president, since Bush won by fewer than 600 votes in Florida.

  The butterfly ballot is only one example of how voting methods can cause confusion. In California, an arrow paper ballot displays two languages, English and either Chinese or Spanish. In Ohio in the 2004 election, Cuyahoga County used punchcard ballots in which the ballot order of names was rotated. Many precincts voted in the same place, and some voters were given a book that did not line up with their ballot or their ballot was processed through the wrong counting machines. This led to between 1,000 and 2,000 votes for Kerry being lost. Yet another example of problems was the 2004 Broward County ballot. It was clearer than the 2000 Florida butterfly ballot, but too much space was left between the candidate's name and where the voter marked his choice.

  For the 2006 election in Mexico, Mebane is using Benford's Law, which examines the frequency of digits, to look at the second digit of vote counts. Results indicate that certain states, such as México and Distrito Federal, show irregularities. The irregularities could be due to votes being swapped or to votes being thrown out as invalid. Mebane used the residual outlier analysis again and found that that one district, Distrito Federal, stood out for its many outliers.

  When the residual vote rates by machine type in the Ohio 2004 election data were studied, much higher rates were found with punchcards than with other machine types. The rate of DREs (direct recording electronic voting machines) fell between those of punchcards and optical scan

machines. Overall, the residual vote rates were not high enough to change the election. A comparison of the 2002 gubernatorial votes to the 2004 presidential votes showed that Kerry had higher turnout in areas where he was strong, and the same was true for Bush. The pattern obtained by this analysis suggested no tampering or switching of votes in Ohio.

Mebane also applied the Benford's Law test of second digits to the Florida 2004 election data. He found that in Miami-Dade County, electronic voting machines do not seem to have been problematic.

Mebane has also studied the problem of auditability. It is true that not all electronic machines get equal numbers or kinds of voters. This can be due to reasons such as crowding. From machine logs, you can tell when each vote was cast. Most machines are used throughout the day, but some are used for only a few hours. This means that all machines are not used randomly, as may have been assumed. This is why precincts satisfy Benford's Law but machines do not. Of course, these records depend on the accuracy of machine logs, that is, to know where the machines were, you need a map of the machines' locations on election day.

When Mebane studied the machine allocation problem in Ohio in 2004, he looked at the correlation with ballots cast per voting machine. The number of ballots cast per machine is lower in areas with higher proportions of African-Americans. Although the ballots were longer in those areas, the difference does not explain the discrepancy in number of registered voters per machine, since in areas where there are higher numbers of African-American voters, polls close later. Mebane's analysis indicates discrimination in the allocation of voting machines.

Mebane uses an assortment of statistical tools that can help assess how voting machines affect voting accuracy, transparency, fraud, etc. Researchers can look at machines in connection with administrative practice and decisions, how they are used in polling places, and how people (both voters and poll workers) respond to machines. Any of these analyses rely on substantial non-quantitative knowledge in addition to the statistics.

Q: You didn't mention exit polls. Do you view them negatively or positively in this?

A: Unless people steal 90% of the votes, exit polls are mostly useless. There was a bias in the 2004 exit polls, a demographic bias and a large sampling error.

Q: About your conclusions: you examined paper and machine ballots?

A: No, I didn't say we should move to paper. I'd recommend optical scan ballots, where the voter gets feedback, the error is reduced, and you have the ballots for recount.

Q: Could you talk more about using Benford's test?

A: It's best to use the second digit, because the test is almost never satisfied by the first digit. If you look at the error rates of when someone means to vote yes and actually votes no and vice versa, then you get this second-digit pattern. If you simulate clumping of votes, you get the second-digit pattern.

## USABILITY

*Summarized by Aaron Burstein*

### *Making Ballot Language Understandable to Voters*

*Sharon J. Laskowski, NIST;*
*Janice (Ginny) Redish, Redish & Associates, Inc.*

The authors examined more than 100 ballots from all 50 states and the District of Columbia, as well as four DRE voting systems, to determine whether their instructions and, in the case of the DREs, system messages conformed with best practices for writing instructions. These best practices were drawn from disciplines such as cognitive psychology, linguistics, and the study of human-computer interaction. Laskowski reported that most, if not all, ballots she and Redish examined violated some best practices. Laskowski highlighted instances of instructions appearing after voting selections; opaque, legalistic language (e.g., "Choose such candidate as you desire"); and instructions whose meaning was obscured by the use of double negatives ("If that oval is not marked, your vote cannot be counted for the write-in candidate"). In addition, some DREs generated system messages that are unlikely to help voters or poll workers understand what problem the DRE system has detected or how to correct it. In addition to recommending that ballot instructions be phrased in clear, direct language and that voters be warned of the consequences of an action before they have an opportunity to take that action, Laskowski outlined several directions for further research. This research should include gaining a better understanding of how voters read a ballot and determining whether voters understand commonly

used ballot terms, such as "cast a ballot," "partisan," "contest," and "race."

A workshop participant asked whether voters actually read ballot instructions; Laskowski replied that she did not know, but that if voters do read instructions, they should be as clear as possible. Another participant asked whether election officials should considering using pictures and images, rather than prose, to convey ballot instructions. Laskowski pointed out that the interpretation of images varies widely with cultural background but that some research into this area might be warranted. Finally, Laskowski stated that some of the guidelines developed—excluding ballot layout guidelines—in this work will be incorporated in Version 2 of the Voluntary Voting System Guidelines (VVSG).

- ***A Comparison of Usability Between Voting Methods***

  *Kristen K. Greene, Michael D. Byrne, and Sarah P. Everett, Rice University*

  Kristen Greene reported the results of the authors' usability studies, measuring efficiency (ballot completion time), accuracy (error rates), and satisfaction (a subjective response), on three traditional voting methods: the open response ballot, the bubble ballot, and the mechanical lever machine. (An open response ballot provides a pair of parentheses within which a voter marks his or her selection but does not indicate what kind of mark the voter should use, whereas a bubble ballot provides an oval that must be darkened to select a candidate.) This study provides a baseline of traditional voting system usability against which electronic voting systems can be compared. A total of 36 subjects participated: 21 female, 15 male; ages 18–56; 23 Rice University undergraduates and 13 subjects from the general population of Houston, Texas. The ballot consisted of 27 races between fictional candidates. Subjects in the study used each ballot type and voted using varying levels of information about the candidates. Greene reported that the three ballot types were generally equally efficient. The ballot types also did not generate statistically different error rates, but the error rate was rather high: 17% of all ballots contained at least one error. Finally, subjects preferred the bubble ballot to the open ballot or lever machines.

  Workshop participants asked several questions about the study's design and the composition of its subjects. Greene stated that subjects from outside Rice were recruited though ads on Craigslist and in the *Houston Chronicle* classified section; the latter subjects displayed an error rate that was significantly higher than the average. The study contained a control for prior voting experience, because elections in the Houston area have previously used punchcard ballots. Finally, in response to questions about differences in voters' incentives in an experimental study versus a real election, Greene acknowledged that voters might take additional care to vote accurately in a real election but noted that neither voters nor test subjects received any tangible incentive to vote accurately. Finally, Greene believes that it is unlikely that governments will devote additional resources to voter conditioning in order to reduce error rates.

- ***The Importance of Usability Testing of Voting Systems***

  *Paul S. Herrnson, University of Maryland; Richard G. Niemi, University of Rochester; Michael J. Hanmer, Georgetown University; Benjamin B. Bederson, University of Maryland; Frederick G. Conrad and Michael Traugott, University of Michigan*

  Paul Herrnson reported the results of usability tests he and his co-workers performed on several electronic voting systems: ES&S Model 100 (paper optical scan ballot), Diebold AccuVote-TS (touchscreen machine with smartcard activation), Avante Vote Trakker (touchscreen with a readable paper printout for verification), Zoomable (a touchscreen prototype developed specifically for this study), Hart Intercivic eSlate (electronic display with a mechanical dial and buttons for navigation and selection), and Nedap LibertyVote (full ballot electronic display with membrane buttons to select candidates). This study was restricted to assessing usability and accuracy in order to develop recommendations for those aspects of electronic voting systems. Herrnson devoted most of his presentation to a field election that involved approximately 1,500 voters but noted that his group's study also included an evaluation of the six voting systems by human-computer interaction experts, a laboratory experiment, and field experiments in Florida and Michigan. Subjects in the field studies were recruited from such diverse locations as inner cities, shopping malls, universities, and business offices. They were asked to complete a ballot with 18 races (more than one selection was allowed in some races), 4 ballot questions, and a write-in option. Study participants indicated a fairly high level of satisfaction with all machines,

with some preference for the Diebold system and significant dissatisfaction with the Hart system. Regarding accuracy, Herrnson reported that study participants successfully cast their ballot for the candidate they wanted 97–98% of the time. Study participants reported that the designs of the ES&S system and the Avante system made it difficult to change their selections. Herrnson stated that few voter characteristics influenced satisfaction, while more education and computer experience, lower age, and greater proficiency with English correlated with fewer help requests and greater accuracy. Finally, Herrnson reported that most field study participants ignored the paper verification features of the ES&S and Avante systems and actually reported a lower level of confidence in those systems than in the Diebold and Zoomable systems.

In response to a question from a workshop participant about accessibility testing, Herrnson said that he and his co-workers had intended to study this aspect of voting systems but lost the part of their budget that was allocated for doing so. Herrnson also noted that his team did not have access to scanners for optical ballots; the researchers had to tally those ballots by hand.

## TECHNOLOGIES

*Summarized by Dan Sandler*

### *Secure Data Export and Auditing Using Data Diodes*

*Douglas W. Jones and Tom C. Bowersox, The University of Iowa*

In order to be communicated to the public, election results must be moved from secure tabulation facilities to public networks. Current best practices involve convoluted chains of dissimilar and obscure computer networks, or physically transported USB storage devices. However complex this chain of networks or disk swaps, each link is bidirectional, so unauthorized communication from the public into the secure inner network is possible.

To create a truly secure transmission system, the authors have devised a data diode, a one-way optical communications medium. What distinguishes the data diode from previous similar approaches is its extreme simplicity: it uses no black boxes or even transistors, so its circuits can be understood and directly inspected. Comprehensive documentation describes the purpose of each component and each trace in the system; the authors call upon all designers of ostensibly verifiable components to do likewise.

A question was asked about timing channels; clearly the diode does not hinder these, and our best tool remains scrupulous analysis of source code on the transmitting side (including deep examination of the serial hardware). Any access to real-time clocks is a red flag. Other measures such as a Faraday cage around the entire tabulation room were proposed by the audience. A pointed question called the big picture "hopeless" even if the diode is a localized success. Jones stressed that the focus of this work is specifically to eliminate the air gap in data transmission, a place where jurisdictions currently make very bad mistakes. By solving this problem we force attackers to resort to other, more challenging attacks.

### *Simple Verifiable Elections*

*Josh Benaloh, Microsoft Research*

True voter verifiability: My vote and *all* other votes are cast as intended and counted as cast. VVPAT (Voter *Verified* Paper Audit Trail) in practice comes nowhere near this goal, but misleadingly implies that it does. We can achieve the goal with complex crypto, but can we achieve it in a way that is understandable and usable by typical voters? Obviously, a completely transparent election—for example, votes posted on a public Web site—achieves this goal, but at the cost of secrecy.

A cryptographic voting system that is trustable and secret should be conceptually simple and require no more of voters than current DREs do. Such a system allows voters to cast encrypted ballots and then verify that those encrypted ballots were tallied correctly (e.g., using reencryption mix nets). When encrypting ballots with potentially untrusted devices, we might use "unstructured auditing," that is, in advance of the election we might allow some voters to create an arbitrary number of encrypted ballots with a device that might be vulnerable. The voter can then choose either to cast each ballot or take it home to check its encryption. A tiny fraction of voters choosing to undertake this audit should detect even a 1% rate of defective encrypted ballots.

Question: With this system I can verify that my own vote was cast and counted correctly, but not others? Answer: You do not know how others voted, but you can still verify that all others were counted correctly.

### *Prerendered User Interfaces for Higher-Assurance Electronic Voting*

*Ka-Ping Yee, David Wagner, and Marti Hearst, University of California, Berkeley; Steven M. Bellovin, Columbia University*

Ping Yee offered a voting machine design in which almost all of the user interface is prerendered long before election day. This design helps jump a number of hurdles facing voting

machine vendors wishing to develop secure systems. The first is accessibility vs. security: making an accessible voting system requires a lot of potentially faulty user-interface code. By prerendering entire ballots we can remove a lot of this UI code from the trusted voting machine, decoupling UI design from security. Anyone could download a prerendered ballot and try it at home, for education or practice or to verify its correctness.

The second issue is that of proprietary code. Vendors would prefer not to disclose code. By reducing the size of the security kernel, vendors can get away with disclosing less. Third, the size of the code base directly affects verification time and complexity; a smaller security kernel is clearly a win here. Finally, vendors worry about the constantly changing requirements for voting machines and the impact on the code base, which must be reverified for each change. The authors argue that a great many of such changes occur in the ballot-design phase of preparing an election, which in their design is removed from the trusted security kernel. The goal is to reduce by an order of magnitude the voting-specific trusted software, with similar or better usability than current systems. The authors' solution consisted of 293 lines of Python and a few libraries.

A member of the audience expressed concern that usability testing isn't being substantially improved by rendering ballots earlier. Ping replied that official usability testing is still essential, but is no longer the last word on the matter, since any constituent is able to download and examine the ballot ahead of time. While it doesn't reproduce the experience of using the voting machine, publishing ballot pictures does allow anyone to vet the interaction. In response to another question, Ping explained that they don't currently plan to apply his techniques to paper (e.g., opscan) ballots. Another participant suggested that the authors investigate usability studies of QWERTY (used in the prototype for write-ins) with other free text-input mechanisms. Finally, Ping reassured a questioner that candidate rotation, i.e., shuffling, is possible with their system by prerendering all the permutations and including them with the final ballot.

■ ***Ballot Casting Assurance***

*Ben Adida, MIT; C. Andrew Neff, VoteHere*

Ben Adida began by saying that voters will, or should, always have concerns about the correctness of voting machines until we offer them end-to-end, voter-centric verification. Voters should have a reasonable assurance that their votes are cast as intended, counted as cast, and not susceptible to coercion or purchase. This talk addresses the cast-as-intended problem, in which we attempt to safeguard the voter's intent until it reaches the ballot box. VVPAT systems address a portion of the chain-of-custody problem—they allow us to ignore the correctness and correct deployment of the voting machine code—but they do not guarantee that results cannot be modified or that they are stored and transported safely.

In VVPAT terms, Ballot Casting Assurance (BCA) means that ballots are cast as intended and the chain of custody is perfect. Such a system might force the voter to revote until the ballot is verified to be acceptable and then give her an authentic receipt that could later be used as evidence in a challenge of count accuracy. Invalid receipts would signal the presence of faulty or malicious voting equipment. The Mark-Pledge and Punchscan systems follow this model. Finally, it is not enough merely to detect errors; we must also supply solid policies for error recovery. The voter's hand-off of the ballot must not be our last opportunity to deal with errors. As David Dill has said, "The difference between using computers for voting and for flying airplanes is that you know when the airplane crashes."

Audience questions prompted discussion of the usability of secure election receipts, especially for large contests. Many options are available to address this particular problem, but only usability testing will tell us for sure which work best for voters. The threat model of the system was clarified: the described systems are intended to detect any malicious software in the voting stack.

## POLICY & PRACTICE

*Summarized by Ka-Ping Yee*

■ ***Transparency and Access to Source Code in Electronic Voting***

*Joseph Lorenzo Hall, University of California, Berkeley*

Transparency and the election process are the foundations of a representative democracy. Hall's definition of "transparency" has four parts: accountability, public oversight, comprehension, and access to the entire process. "Open source" can refer to the open source license or to the development model. Source code can also be disclosed even if the disclosure doesn't include all the components of the official Open Source Definition. Though computer scientists often say that all voting code should simply be made open source, the issue is

more complex than that: it has both positive and negative effects on security and on the market.

Source availability offers several benefits: more people can examine the code; you can build the code yourself and debug it; you can use automated tools to evaluate it. However, software alone is not enough. For a full evaluation you need access to the complete system in its running environment. Some states are starting to require code escrow and disclosure. Open source also brings risks. It exposes vulnerabilities to the public, and it would require a process for handling flaws discovered just before an election.

Barriers to disclosing source code for voting technology include: (1) regulations require system recertification whenever code changes; (2) certification and contractual performance bonds are expensive; and (3) to field a product, you need more than just code development. It remains an open question how we can level the playing field for open source or disclosed source. As an incentive, the government might offer a prize in a Grand Challenge to develop an open source voting system, subject to some requirements. It may be very difficult for vendors to move to a disclosed source regime, because their code wasn't designed to be exposed; it may contain patented work or work improperly copied from other sources, for example.

Question: If you designed your system not thinking about it being opened, what will you do when it finally leaks? Even when there are strong controls on source code access, it seems often to be leaked or reverse-engineered. Answer: Maybe we need to put vendors on notice that you need to design your code as if you have nothing to hide. Maybe it's time to start now. An audience member commented that with regard to foundations for transparency in a representative democracy, we might look at Arrow's impossibility theorem: in order to verify the conditions of the theorem, such as that the decision is not imposed or that the decision responds positively to changes in individual preference, you would need transparency. Another participant commmented that Arrow's theorem is about the process of vote tallying, but you could disclose the tallying software without disclosing the vote selection software.

Question: What do you think would happen if federal legislation immediately mandated software disclosure? Answer: Because vendors compete on razor-thin margins, you may see an exodus. But some vendors are more confident about the quality of their code than others. I'm not really sure what would happen.

- ***A Critical Analysis of the Council of Europe Recommendations on E-Voting***

  *Margaret McGaley and J. Paul Gibson, NUI Maynooth, Ireland*

  McGaley explained that the Council of Europe, CoE, is an organization of 46 member states and is not directly connected with the EU. In 2003 the CoE created a committee to develop legal, operational, and technical standards for electronic voting. E-voting was first deployed in Europe in 1982 (in the Netherlands) and then in 1991 (in Belgium) and has since been tested in the U.K., Italy, Spain, and Ireland. The U.K. and Ireland are pulling back from their more ambitious plans for various reasons, including some detected fraud in postal voting.

  The U.S. standards effort is older. The first FEC (Federal Election Commission) standards were produced in 1990, whereas the CoE document is only two years old. The U.S. standards are nominally voluntary but in many states are legally required. In Europe, only Belgium appears to be using the CoE standards, which are shorter and less detailed than the FEC standards.

  The authors evaluated the CoE standards from a software engineering perspective: they examined consistency, completeness, scope, over/underspecification, redundancy, maintainability, and extensibility, Many problems were uncovered: Some of the standards are vague, ill-defined, or nonsensical, although it is conceivable that better systems might fail to meet these standards while worse systems might pass.

  The authors propose a restructuring of the standards, categorizing them according to the five basic rights identified in the original standard: that they ensure universal, equal, free, secret, and direct suffrage. Organizing the standards in this fashion prevents inconsistency and redundancy, maximizes coverage, and makes them easier to understand and use. In their proposed restructuring, some of the standards are merged, some are revised or omitted, and some additional standards have been added.

  Question: You mentioned bug-tracking software in your proposed standard. Were you thinking about soliciting comments during the use of a voting system and incorporating the changes during a further development process? Answer: What we had in mind is that each bug would have an identifier and would be traceable as to how it was resolved or not resolved. The system purchased by the Irish government didn't have any sort of bug-tracking system, so after a

problem was reported, it was hard to trace. Question: Is anybody at the CoE listening to your recommendations? Answer: They are: one of the members read our paper and was very interested in it. Question: Does the CoE ever solicit input from nonmember nations or international organizations? Answer: Yes. In fact, Canada is a regular participant!

- ***An Examination of Vote Verification Technologies: Findings and Experiences from the Maryland Study***

  *Alan T. Sherman, Aryya Gangopadhyay, Stephen H. Holden, George Karabatis, A. Gunes Koru, Chris M. Law, Donald F. Norris, John Pinkston, Andrew Sears, and Dongsong Zhang, University of Maryland, Baltimore County*

  Sherman explained what his group found when they evaluated four vote verification products: a Diebold VVPAT, an MIT audio system developed by Ted Selker, a software system called Scytl Pnyx.DRE, and the VoteHere system based on cryptographic receipts. By 2007 Maryland will have spent $96 million on Diebold systems. The authors believe that governments should spend some fraction—even if only 2 percent—of that money on voting system research. Their study looked only at how vote verification products worked with the Diebold voting system, not at whether the voting system as a whole is secure. Adding verification to the system would be challenging, since it would add complexity and would require that Diebold revise their software.

  The authors evaluated each of the verification products in terms of reliability, functional completeness, accessibility, data management, election integrity, implementation difficulty, and impact on voters and procedures. Each product could probably improve the situation somewhat, but none is a fully ready product. For example, the Diebold VVPAT can't be used by blind voters, and the MIT-Selker audio system can't be used by deaf voters. Also, integration with the DRE machine can be complicated; indeed, the Scytl Pnyx.DRE system can cause the DRE to fail. The VoteHere cryptographic system provides strong election integrity and is implemented in high-quality open source software. However, it may be more difficult for the user. Parallel testing, a powerful technique, was found it to be in some ways better than these vote verification products.

  Question: I don't share your confidence in parallel testing. It doesn't seem particularly difficult for malware to beat parallel testing, even if it's conducted fairly carefully. Answer: The easiest way to subvert parallel testing is to load the wrong software onto all the machines and then signal the machines that are being tested to operate correctly. I don't mean to imply that parallel testing is perfect, but I do believe it meaningfully raises the bar by addressing the thread of systemic failure. Question: I'm surprised you rated Scytl higher in terms of election integrity than VVPAT. Could you elaborate on why? Answer: Scytl uses cryptography to protect the information in more places, as compared to the chain of custody issues of a VVPAT.