Towards a Privacy Measurement Criterion for Voting Systems*

Lillie Coney
EPIC
coney@epic.org

Joseph L. Hall
U.C. Berkeley
joehall@berkeley.edu

Poorvi L. Vora
George Washington Univ.
poorvi@gwu.edu

David Wagner
U.C. Berkeley
daw@cs.berkeley.edu

ABSTRACT

This paper provides a definition for perfect privacy in electronic and other voting systems, and an entropy-based criterion to measure the deviation from perfect privacy. Its use is illustrated with examples.

1. INTRODUCTION

The voting process is just one more in the list of valuable community activities that can benefit from the strengths of today's information technology. The use of computers does, however, introduce problems of the kind not seen in paper-based voting systems. A taxonomy/classification of electronic voting systems—based on their satisfaction of a list of desirable properties—is hence urgently needed so that (a) standards bodies may use it to develop a performance rating standard, and (b) technical aspects of the debate can be disentangled from political and emotional aspects, as well as those of pragmatic expedience [1].

This paper addresses the problem of defining and measuring privacy in voting systems. A quantification of the extent of privacy loss in different voting systems would enable several meaningful comparisons. For instance, if we reveal the entire contents of all cast ballots, does this reveal more information on how Alice voted than if we only release vote totals? If so, how can we measure the magnitude of this privacy loss? As another example, suppose we are comparing a precinct optical scan system—where all voters in the precinct cast their ballots into a single ballot box—to a system where the precinct has five touchscreen machines and each machine maintains its own virtual ballot box. In the latter approach, ballots are separated into five small pools, instead of one large pool of ballots. Does this provide less privacy? If so, how much less?

We seek to analyze both the extent of privacy invasion possible with and without the voter's collusion. Vote buying and voter coercion can result when systems allow the voter to provide proof of how she voted. Thus, we can try to measure the amount of privacy lost if the voter is colluding with a vote-buyer, versus the amount of privacy lost for honest voters who prefer to keep their vote secret.

2. PERFECT PRIVACY

There are many sources of information on how a voter might vote. For instance, race, geographical location, and economic status are known to be correlated with vote choices. A voter may herself reveal information that tends to predict her vote, for instance by publicly contributing to a party's campaign fund or by buttons and bumper stickers advertising a particular candidate. An adversary might be able to use these sources to gain partial information on how the voter is likely to vote.

The voting process and system cannot prevent information leakage from these sources. It can, however, provide privacy to the voter by not leaking any further information regarding her vote. Thus, we consider a voting system to provide perfect privacy to the voter if the voting system itself reveals no further information on the voter's vote.

Definition 1. An election system is *perfectly private* if an adversary's information about a voter's ballot choice(s), as obtained through the election technology and process/procedures, is not affected by the actual vote cast by the voter on election day.

In other words, a voting system/process provides complete privacy to the voter if anything the adversary could have learned about the voter's choices (e.g., via information leaked by the system), could also have been learned even if the voter had cast some other vote.

This definition can be made more precise as follows. Let the random variable V denote the voter's vote (as actually cast), S denote the information through sources other than the voting system (e.g., geographic location), and E denote the information revealed to the adversary by the voting system and process. In general, these r.v.'s may be correlated in complex ways. To simplify notation, let p_X denote the probability distribution of a r.v. X.

Definition 2. An election system is *perfectly private* if V is conditionally independent of E after conditioning on S, i.e., $p_{V|S}(v;s) = p_{V|S,E}(v;s,e)$ for all v, s, e.

We assume E includes all information (i) through/due to the voting technology and poll place procedures, (ii) available to poll workers, election officials, and other insiders (not merely what is visible to outsiders), and (iii) that is stored in any permanent form. However, we assume the adversary is not coercing, or colluding with, the voter in any way, and we assume that the voter prefers to keep her vote secret for the purposes of Definition 2.

Example 1. We have two candidates, a and b. Before election day, the adversary estimates Alice is equally likely to vote for both candidates: $p_{V|S}(a;s) = p_{V|S}(b;s) = \frac{1}{2}$. The voting system reveals partial information e so that $p_{V|S,E}(a;s,e) = \frac{2}{3}$. Then the adversary's estimate of Alice's vote has improved, and

*This work is being carried out in the Privacy Measurement Criteria Working Group of the Voting Systems Performance Rating (VSPR) project.

the election system is not perfectly private. Note the connection to Shannon's definition of perfect secrecy for a cryptographic system.

3. BALLOT SECRECY

To prevent voter coercion and vote buying, we may also wish to ensure that the voter herself cannot prove how she voted. This property is termed *ballot secrecy*.

Definition 3. An election system has *perfect ballot secrecy* if it is perfectly private even when the voter is in collusion with the adversary and even if the voter wishes to prove how she voted to the adversary.

4. MEASURES OF IMPERFECTION

To measure the privacy of a voting system, one would wish to measure how much the voting system deviates from perfect privacy. In particular, we seek a numerical measure of how much $p_{V|S,E}$ differs from $p_{V|S}$ (Definition 2). We propose a definition that measures the reduction in the "uncertainty" in the vote due to the gained knowledge leaked by the system and process. The reduction in uncertainty will typically depend on both the distribution $p_{V|S}$—i.e., how well the vote may be estimated without information leaked by the voting system and process—and $p_{V|S,E}$—i.e., the form of the leakage due to the voting system. We focus on the latter. In particular, we would not wish to characterize a system as "good" simply because the initial uncertainty in the vote is very small and thus knowledge of E has little or no effect. Also, it is hard to predict how the system will be deployed, so we may not have advance knowledge of $p_{S|V}$ or p_V. Therefore, we propose that our definition consider the worst-case uncertainty reduction over all possible "prior" distributions $p_{V,S}$.

Definition 4. The *amount of privacy loss* of a voting system and process is the maximum reduction in uncertainty of a voter's vote due to information revealed by the election system and process.

We propose that Shannon entropy be used to measure this difference. Entropy is a mathematical measure of the uncertainty in a r.v. X, defined by $\mathcal{H}(X) = -\sum_x p_X(x) \lg p_X(x)$. Roughly speaking, the entropy $\mathcal{H}(X)$ is the minimum number of bits required, on average, to represent variable X.

This allows a more precise definition of privacy loss:

Definition 5. The *amount of privacy loss*, $\mathcal{L}$, of a voting system and process is

$$\mathcal{L} = \max_{p_{V,S}} \mathcal{H}(V|S) - \mathcal{H}(V|S,E),$$

where $p_{E|V}$ is held fixed and $p_{V,S}$ varies.

It can be shown that $\mathcal{H}(V|S) - \mathcal{H}(V|S,E)$ is always non-negative; this quantity is known as the *conditional mutual information* between the vote and the information leaked by the voting system and process, conditioned on information from other sources. To measure the amount of privacy left after the use of a voting system to cast, say, N votes, one might perform a worst-case analysis: $\mathcal{H}(V|S,E) \geq \mathcal{H}(V|S) - N \times \mathcal{L}$. Because it will be difficult for an adversary to reduce the entropy by the maximum amount for each vote, this bound is far from tight.

The distribution $p_{E|V}$ is very important: it characterizes the way that the voting system leaks information about the voter's vote. We assume that E depends only on the value of V, so that $V \rightarrow E$ is a Markov chain, i.e., $p_{E|V} = p_{E|V,S}$. We make a similar assumption about S.

The voting system may be considered a communication channel, with the vote as input and the adversary as receiver. Thus $p_{E|V}$ represents the "forward channel" characterization of this channel, and $\mathcal{L}$ represents the capacity of the voting communication channel—a measure of its ability to carry information. Also, $\mathcal{L}$ is zero if and only if $p_{V|S} = p_{V|S,E}$, which means that the privacy loss is zero if and only if the system is perfectly private (see Definition 2).

Example 2. During the Nov. 2004 elections, Nevada used Sequoia AVC Edge machines with VVPAT printers. These machines allow a voter to begin voting while the previous voter's VVPAT record was still scrolling up onto the take-up reel for storage. For simplicity, assume the choices were between candidates a and b, as in Example 1. Then E, the information obtainable by peeking at the VVPAT, takes on a value from $\{a, b, ?\}$, where $?$ represents the event that the VVPAT has completely scrolled. Based on an analysis of the time between voters and the time it takes for the VVPAT completely scroll, one might determine that about a fraction ϵ of the votes cast would be revealed correctly to the next voter, and that no information at all would be revealed about the other $1 - \epsilon$ fraction. This is a binary erasure channel, with $p_{E|V}(v; v) = \epsilon$ for $v \in \{a, b\}$ and $p_{E|V}(?; v) = 1 - \epsilon$.

The amount of privacy loss may be computed as follows. A brief calculation shows $\mathcal{H}(V|E) = (1 - \epsilon) \times \mathcal{H}(V)$. Also, $\mathcal{H}(V)$ is maximized if $p_V(a) = p_V(b) = \frac{1}{2}$. Hence, $\mathcal{L} = \max_{p_V} \mathcal{H}(V) - (1 - \epsilon)\mathcal{H}(V) = \epsilon$. This agrees with intuition: the privacy loss is proportional to the chance that the next voter can see the previous voter's VVPAT record.

5. CONCLUSIONS

This abstract sketches the beginnings of an entropy-based approach to the definition and measurement of the privacy of voting systems. A number of questions remain. How large may $\mathcal{L}$ be for good voting systems? How easy or difficult will it be to use this to measure the privacy of real voting systems? How effective will it be? These are questions our research—which is currently in its very early stages—hopes to address.

6. REFERENCES

1. P.Vora, B.Adida, R.Bucholz, D.Chaum, D.Dill, D.Jefferson, D.W.Jones, W.Lattin, A.D.Rubin, M.I.Shamos, and M.Yung. Evaluation of voting systems. *Communications of the ACM*, 47(11):144, November 2004.