

APPROXIMATE MEASUREMENT OF VOTER PRIVACY LOSS IN AN ELECTION WITH PRECINCT REPORTS

CHRISTOPHER CRUTCHFIELD, DAVID MOLNAR, AND DAVID TURNER

ABSTRACT. The California election process publishes tallies for each precinct as part of each county’s Statement of Vote. We take a data set from two California counties in the November 2004 election and use it to measure an approximate *voter privacy loss* : how much information is leaked about a voter’s vote by publishing precinct tallies. Our starting point is the privacy measurement framework of Coney et al. [6]. We summarize the framework, then we discuss issues with attempting to use the framework directly. We introduce simplifying assumptions and argue that these assumptions still yield meaningful results. We then show how to measure approximate voter privacy loss given real election data. We also show how to incorporate the effect of pre-election polls into measuring privacy loss.

1. INTRODUCTION

Reporting precinct-level tallies after an election is a widespread electoral practice. We show, using data from the November 2004 elections, that reporting precinct-level tallies leaks information about how voters voted. In particular, we point out rare — but real — cases where precinct tallies, combined with other public information, unambiguously identify a voter’s choice in the election. Inspired by these cases, we generalize the framework of Coney et al. to *quantify* privacy lost by revealing precinct-level tallies in general [6]. We report on experience with using the framework, compare the loss to the loss from reporting solely on a county-by-county basis, and critically evaluate our notion of privacy loss.

Our study focuses on data from two Californian counties: San Francisco and Santa Cruz. We chose these counties because they report precinct-level tallies and they have Excel spreadsheets of their data available online. Figure 1 shows precincts in these counties where each and every voter made the same choice for President. In the case of Santa Cruz, the “800” prefix indicates absentee ballots for the precinct number following, for example “8004324” refers to absentee ballots for precinct 4324. In the case of San Francisco, these precincts are marked as “mail ballot” precincts. If we could obtain the list of registered voters in these precincts and determine which voters participated in the November 2004 election, then we would know these voters’ choices exactly.

As it turns out, exactly these data and more are collected by both the Santa Cruz and San Francisco registrar of voters. In Figure 2, we give a partial list of the data available in each county’s voter file. California law states that this data is confidential information, but California law also sets out specific conditions under which this information may be released to third parties. Four major categories of users exist that may obtain data, after signing an agreement under penalty of perjury promising to keep the data confidential : journalists, scholars, campaigns, and government agencies [10]. In particular, political campaigns regularly obtain a county’s voter file to target likely supporters in an election.

Date: June 3, 2006.

County	Precinct Number	Bush	Kerry	3rd-party
Santa Cruz	8004324	0	6	0
Santa Cruz	8005000	0	3	0
Santa Cruz	8005009	0	1	0
San Francisco	3938 (V)	0	6	0
San Francisco	3101 (V)	0	1	0
San Francisco	2502 (V)	0	4	0
San Francisco	2219 (V)	0	12	0

FIGURE 1. Selected precincts from the Statement of Vote for the November 2, 2004 election. Each precinct has a 100% vote for its candidate of choice.

Name
 Phone number
 Home address
 Precinct number
 Voted in previous election
 Voted absentee or not
 Party registration

FIGURE 2. Partial listing of data available in the San Francisco and Santa Cruz voter file.

We have checked that one can obtain data from both the Santa Cruz and San Francisco voter file that links voter names with precinct numbers. Obtaining this data requires signing a confidentiality agreement, a nominal fee, and a declaration that the data is for research purposes. The 2004 Voting Privacy Task Force reports on other procedures surrounding the release of voter file information, and makes recommendations aimed at limiting the unnecessary release of this data [10].

The fact remains, however, that the intended political uses of the data need the list of voters in each precinct. If both this information and the precinct tally is available, then the result is catastrophic privacy loss in the rare but real cases we have shown. These cases are not just an academic concern. The Riverside County (California) registrar of voters refused to post precinct tallies and explicitly cited potential privacy loss as a reason, even though such practice is ordered by the California Secretary of State and enshrined in California law [4].

Reporting precinct tallies, on the other hand, has a public benefit. If the precinct tally is public, then anyone can watch the counting of each precinct's ballots. Furthermore, some of today's audit procedures rely on observing an individual voting machine's tallies, which in turn reveals the precinct tallies. Campaigns also depend on precinct tallies to focus their energies in an election. Also, as mentioned, California law and policy requires that precinct tallies be disclosed; several states have similar laws. In general, public precinct tallies enhance the transparency and ease of audit of an election.

Therefore, there is a trade-off between voter privacy and the public benefit from revealing precinct-level tallies. To evaluate this trade-off and compare different strategies for preserving privacy, we need to go beyond the pathological cases shown above. It is clear that unanimous outcomes lose privacy. How do we reason about non-unanimous outcomes? Put another way, what happens to voter privacy in precincts that are not 100% for a single candidate? There are a wide variety of

changes one could consider for election practices, but picking between them requires a principled way to measure privacy loss. Our study addresses this question by starting with a quantitative theory of privacy loss due to Coney et al. [6].

We make three main contributions. First, we identify the problem of privacy loss due to precinct tally reports and the resulting tension between voter privacy and election transparency. Second, we report experience with applying the framework of Coney et al. to quantify privacy loss from precinct tallies. In particular, we introduce generalizations of their method and simplifying assumptions in Section 2 that allow us to compute privacy loss from real election data in Section 3.1. We also show how additional information about voter preferences prior to voting, such as polls, can affect privacy loss.

Our work does not answer the question “should counties release precinct tallies?” What we do, instead, is explore one framework for measuring the privacy impact of releasing this information. Indeed, as noted above, several states have laws that mandate releasing precinct tallies.

In Section 7 we step back and consider whether our results indicate that the framework we use is the “right” one for this task. Finding the right framework will allow others to compare different approaches for dealing with privacy loss from revealing precinct tallies. For example, one possible approach, suggested by Hall, is to consolidate reports for several precincts after the election in cases where one precinct has a unanimous outcome [9]. This approach rules out the pathological cases we have shown, but we cannot fully compare it to other approaches without a better understanding of privacy loss.

Related Work. Coney et al. introduced the framework we use for privacy measurement [6]. Our problem can be seen as a special case of the general *census problem*, in which a database administrator must decide what information to release to strike a balance between privacy and utility. Chawla et al. suggest formal definitions of privacy in these cases and survey a wide variety of work in the statistics and databases communities[3]. Looking at their definitions to see if they can be applied to our problem would be interesting future work, but we chose to focus on an approach already proposed for voting. Pelote et al. report on privacy issues concerning data in a county voter file, but they focus on address and phone number information and they do not discuss information released as part of the Statement of Vote [10].

2. PRELIMINARIES

We first summarize the model of Coney et al. Let the random variable V denote the distribution of a voter’s vote. Let S denote the information available to an adversary through sources other than the voting system, such as geographic location or party affiliation. Let E denote the information revealed to the adversary by the voting system; in our study, E will be the precinct tallies released at the end of the election. We denote the probability distribution of a random variable X by p_X .

Recall that the Shannon entropy $H(X)$ of a random variable X is defined as

$$H(X) = - \sum_x p_X(x) \log p_X(x).$$

The *amount of privacy loss*, \mathcal{L} , of a voting system is

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

Informally, the privacy loss \mathcal{L} of a voting system quantifies “how much” an adversary learns from the information E revealed by the voting system. In our case, we would like to measure \mathcal{L} in a slightly different context: instead of looking at a whole voting system, we focus only on the

reporting policy for an election. In particular, we would like to measure the privacy loss \mathcal{L} for a reporting policy that releases precinct tallies after an election. But to do so, we need to look at *aggregate* privacy loss; that is, the total privacy loss across all of the voters. In order to do this, we need to modify the model presented by Coney et al. to reflect this. We consider each voter's vote to be a random variable V_i , and then consider V as the joint distribution over all of the V_i . One critical assumption is that the V_i 's are all independent random variables (which, of course, may not correspond to reality). As before,

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

Unfortunately, directly calculating \mathcal{L} is difficult. The random variables V, E , and S may depend on each other in complicated ways. This makes maximizing over $p_{V,S}$ difficult. Instead, we propose making some simplifying assumptions. Our simplifications will allow us to calculate approximate privacy loss from real election return data.

To begin, we simplify the model by assuming that the adversary is given no information through channels outside of the reporting policy; that is, the effect of S is negligible. In addition, the adversary is given no information through the reporting policy, except for a list of vote tallies per precinct after the elections (this is the effect of E). Finally, we assume each voter votes independently. Therefore we want to compute

$$\mathcal{L} = \max_{p_V} (\mathcal{H}(V) - \mathcal{H}(V|E)).$$

Later on we will relax some of these assumptions to get a more realistic model.

3. PRECINCT TALLIES FOR TWO-CANDIDATE ELECTIONS

Let us consider the simple model of an election with two candidates. Let V_i correspond to the indicator variable for voter i 's vote. Recall that we assumed each voter votes independently, so V_i is 1 with probability p_i , and 0 with probability $1 - p_i$. Let V be the joint distribution of all the V_i , so $V = (V_1, V_2, \dots, V_n)$. Now define the random variable $X = V_1 + V_2 + \dots + V_n$.

We notice that E is just an observation of a tally, so our calculation comes out to be

$$\mathcal{L} = \max_{p_V} (\mathcal{H}(V) - \mathcal{H}(V|X = k)).$$

Maximizing over V is really maximizing over $p_1, p_2, \dots, p_n \in [0, 1]$, since the V_i are independent. Instinctively, one might assume that \mathcal{L} reaches its maximum when $p_1 = p_2 = \dots = p_n = 1/2$, the *uniform case* V_U . However, this is not always true, as in the following example.

Example 3.1. Let $n = 3$ and $k = 2$. In addition, let V be the joint distribution for the uniform case V_U , $p_1 = p_2 = p_3 = 1/2$. Then

$$\begin{aligned}
\mathcal{H}(V) - \mathcal{H}(V|X = 2) &= 3 + p_{V|X=2}(110) \log_2(p_{V|X=2}(110)) \\
&\quad + p_{V|X=2}(101) \log_2(p_{V|X=2}(101)) \\
&\quad + p_{V|X=2}(011) \log_2(p_{V|X=2}(011)) \\
&= 3 + 3 \left(\frac{1}{3} \log_2 \left(\frac{1}{3} \right) \right) \\
&= 3 - \log 3 \\
&\approx 1.415
\end{aligned}$$

Now suppose that V' is the joint distribution for the slightly different case, $p_1 = 0$, $p_2 = p_3 = 1/2$. Then

$$\begin{aligned}
\mathcal{H}(V') - \mathcal{H}(V'|X' = 2) &= 2 + p_{V'|X'=2}(011) \log_2(p_{V'|X'=2}(011)) \\
&= 2
\end{aligned}$$

So it turns out that if the effect of S is negligible, and E is the observation of a tally on V , \mathcal{L} is always maximized for the degenerate case where $n - k$ of the p_i are 0. However, for our purposes, this is not a useful calculation; and, in general, it will not be true to our observations of the real world. Instead, we consider the case where the adversary knows nothing about the distribution V . To model this case, we set V to be the uniform prior V_U . We then define the value \mathcal{L}' as follows:

$$\mathcal{L}' = \mathcal{H}(V_U) - \mathcal{H}(V_U|E) \leq \mathcal{L}$$

The value \mathcal{L}' corresponds to the case where the adversary knows nothing about the voter's prior preferences. In general, $\mathcal{L}' \leq \mathcal{L}$, and therefore provides a lower bound for the true privacy loss.

Suppose that there are m precincts, P_1, P_2, \dots, P_m . Each precinct P_i has n_i voters, among whom k_i voted 1 (and therefore $n_i - k_i$ voted 0). We let X_i denote the joint distribution of votes for the voters in precinct P_i . In addition, we notice that the X_i remain conditionally independent on E , hence $Pr(V_U|E) = Pr(X_1, X_2, \dots, X_m|E) = \prod Pr(X_i|E)$. Therefore all that remains is to compute the distribution $X_i|E$. This is just the uniform distribution on n_i -bit strings containing exactly k_i 1's and $n_i - k_i$ 0's. Therefore

$$\begin{aligned}
\mathcal{L}' &= \mathcal{H}(V_U) - \mathcal{H}(V_U|E) \\
&= n - \mathcal{H}(X_1, X_2, \dots, X_m|E) \\
&= n - \sum_{i=1}^m \mathcal{H}(X_i|E) \\
&= n - \sum_{i=1}^m \left(- \sum_{v \in X_i} p_{X_i|E}(v) \log_2 p_{X_i|E}(v) \right)
\end{aligned}$$

Since each member of the distribution $X_i|E$ occurs with probability $1/\binom{n_i}{k_i}$, we have

$$\begin{aligned} \mathcal{L}' &= n - \sum_{i=1}^m \left(- \sum_{j=1}^{\binom{n_i}{k_i}} \frac{1}{\binom{n_i}{k_i}} \log_2 \frac{1}{\binom{n_i}{k_i}} \right) \\ &= n - \sum_{i=1}^m \left(\sum_{j=1}^{\binom{n_i}{k_i}} \frac{1}{\binom{n_i}{k_i}} \log_2 \binom{n_i}{k_i} \right) \\ &= n - \sum_{i=1}^m \log_2 \binom{n_i}{k_i} \end{aligned}$$

3.1. Application to Real Data. We can now apply our notion of privacy loss to real data. We used precinct tallies from the November 2004 Statement of Vote from San Francisco and Santa Cruz counties [7, 8]. For each county, we computed the privacy loss \mathcal{L}' for two scenarios. In the first scenario, we assume that the tallies for each precinct are given. In the second scenario, we assume that only the county-wide tally is given. Intuitively, the second scenario is “more private,” because it releases less information. In particular, revealing only county-level tallies prevents privacy loss for the rare cases described above.

Precinct tallies			County-Only		
n	\mathcal{L}'	\mathcal{L}'/n	n	\mathcal{L}'	\mathcal{L}'/n
351127	146005	0.415819	351127	132822	0.378273

FIGURE 3. Data from the San Francisco November 2004 Statement of Vote. Reports the privacy loss \mathcal{L}' and average privacy loss for two scenarios: the first where precinct totals are reported, and the second where only the county-wide tally is released.

	Precinct tallies			County-Only		
	n	\mathcal{L}'	\mathcal{L}'/n	n	\mathcal{L}'	\mathcal{L}'/n
All data	119456	28990.5	0.242688	119456	21783.6	0.182357
No absentee	73838	20896.9	0.28301	73838	16820.5	0.227802
Only absentee	45618	8093.61	0.177422	45618	5514.57	0.120886

FIGURE 4. Data from the Santa Cruz November 2004 Statement of Vote.

In each case, we report n , the number of voters, and \mathcal{L}' , the approximate privacy loss calculated as above. We also report the *average privacy loss*, which we define to be \mathcal{L}'/n . The results for San Francisco are in Figure 3.

For Santa Cruz, we noticed that the value of \mathcal{L}' changed significantly depending on whether absentee ballots were included. Therefore, when we report the results in Figure 4, we report two more scenarios. In one, we consider precinct-level tallies without absentee ballots; we can do this because the Santa Cruz statement of vote reports absentee ballots as precincts starting with “800.” In the other, we consider only absentee ballots.

3.2. Discussion. Our first observation is that revealing precinct tallies leads to a large average privacy loss in San Francisco, with a value of 0.41. Surprisingly, revealing the county tally alone leaks almost as much privacy, with a value of 0.37. We attribute this to the fact that we used a uniform prior, while San Francisco voted overwhelmingly for Kerry in the 2004 election. This tells us that our uniform prior may be “too pessimistic” for the case of San Francisco; it does not take political trends of the city into account. As a contrast, we see that the absentee-only case in Santa Cruz has a relatively low average privacy loss, with a value of 0.16. In Section 5 we discuss one way to pick a more realistic prior.

4. ELECTIONS WITH MULTIPLE CANDIDATES

In this section we examine how to analyze elections where there are more than two candidates. In general, this is often the case. Besides the Democrat and Republican party candidates, there are also Libertarian party, Green party, Constitution party, Peace and Freedom party, as well as write-in candidates. Inclusion of this data into our calculations might affect the amount of privacy loss that we see, especially considering that some of these third parties occasionally receive a fairly significant portion of the vote (for example, Ross Perot in 1992).

Let’s assume that we have ℓ candidates. As before, we let n_i be the total number of votes cast in precinct i . Then $k_{i,j}$ is the number of votes cast for candidate j in precinct i .

$$\begin{aligned}
\mathcal{L}' &= n \log_2 \ell - \sum_{i=1}^m \log_2 \left(\binom{n_i}{k_{i,1}} \cdot \binom{n_i - k_{i,1}}{k_{i,2}} \cdots \binom{k_{i,\ell-1} + k_{i,\ell}}{k_{i,\ell-1}} \right) \\
&= n \log_2 \ell - \sum_{i=1}^m \log_2 \left(\frac{n_i!}{k_{i,1}!(n_i - k_{i,1})!} \cdot \frac{(n_i - k_{i,1})!}{k_{i,2}!(n_i - k_{i,1} - k_{i,2})!} \cdots \frac{(k_{i,\ell-1} + k_{i,\ell})!}{k_{i,\ell-1}!k_{i,\ell}!} \right) \\
&= n \log_2 \ell - \sum_{i=1}^m \log_2 \left(\frac{n_i!}{k_{i,1}!k_{i,2}! \cdots k_{i,\ell}!} \right) \\
&= n \log_2 \ell - \sum_{i=1}^m \left(\log_2(n_i!) - \sum_{j=1}^{\ell} \log_2(k_{i,j}!) \right)
\end{aligned}$$

Precinct tallies			County-Only		
n	\mathcal{L}'	$\mathcal{L}'/n \log_2(\ell)$	n	\mathcal{L}'	$\mathcal{L}'/n \log_2(\ell)$
358081	751532	0.747599	358081	722505	0.718724

FIGURE 5. Data from the San Francisco November 2004 Statement of Vote. Note that $\ell = 7$, as there are 7 candidates.

Precinct tallies			County-Only		
n	\mathcal{L}'	$\mathcal{L}'/n \log_2(\ell)$	n	\mathcal{L}'	$\mathcal{L}'/n \log_2(\ell)$
121733	206942	0.657637	121733	196368	0.624035

FIGURE 6. Data from the Santa Cruz November 2004 Statement of Vote. Note that $\ell = 6$. Number of write-ins were not reported.

Note that when considering multiple candidates, our average privacy loss is now $\mathcal{L}'/n \log_2(\ell)$, since each voter now contributes $\log_2(\ell)$ bits instead of $\log_2(2) = 1$.

4.1. Drawbacks. The choice of prior for races with multiple parties is a further problem. In U.S. politics, most races are dominated by the two major parties, with third parties receiving a small proportion of the vote, although as noted some exceptions exist. Duverger’s law explains that first-past-the-post plurality voting naturally tend toward two major candidates [1]. Therefore, in most cases, a uniform prior will overestimate the amount of privacy loss.

For example, if we have three parties, Democrat, Green, and Republican, the uniform prior states that the Green party candidate has about a 33 percent chance of winning. If the Green candidate in fact receives 2.74 percent of the vote, as Nader did in 2000, then our calculations will report a large privacy loss [5]. A more “realistic” privacy analysis would take into account prevailing political trends available before the election. In the next section, we show how our side information variable S can encode such information.

5. INCLUDING A REALISTIC PRIOR DISTRIBUTION

In the previous sections, we concerned ourselves with the case where S reveals no information about the prior distribution. However, as we’ve shown, this is not a realistic way to model what goes on in the real world. Oftentimes there is a large amount of data released about how a particular voter might vote prior to the election — political party affiliation, or even polling data. How do we adjust our model to take into account the effect of S ?

5.1. Polls. One condition we might consider is that an adversary might have access to some polling data prior to the election. For example, he might know that with high certainty that in San Francisco County, 85 percent of the people will vote for Kerry, while 15 percent will vote for Bush. This then biases the prior distribution. We can model this in the following way: our indicator variable V_i is now 1 with probability 0.72 and 0 with probability 0.28. Therefore our privacy loss is redefined as

$$\mathcal{L}' = n\mathcal{H}(V_i) - \sum_{i=1}^m \left(\log_2(n_i!) - \sum_{j=1}^{\ell} \log_2(k_{i,j}!) \right).$$

In order to simulate the release of poll data prior to the election, we base the distribution of V_i on the final county-wide tally released. For the San Francisco computations, we used the tally: (Peroutka, Peltier, Kerry, Cobb, Bush, Badnarik, Write-In) = (380, 1167, 296772, 1854, 54355, 1401, 2152). For the Santa Cruz computations, we used the tally: (Peroutka, Peltier, Kerry, Cobb, Bush, Badnarik) = (327, 404, 89102, 782, 30354, 764).

County	Precinct tallies			County-Only		
	n	\mathcal{L}'	\mathcal{L}'/n	n	\mathcal{L}'	\mathcal{L}'/n
San Francisco	351127	17538.5	0.0499493	351127	9.06949	0.0000258297
Santa Cruz	119456	7215.45	0.0604026	119456	8.55907	0.0000716504

FIGURE 7. Non-uniform prior with two candidates (Bush and Kerry). Data from the San Francisco and Santa Cruz November 2004 Statement of Vote.

County	ℓ	Precinct tallies			County-Only		
		n	\mathcal{L}'	$\mathcal{L}'/n \log_2(\ell)$	n	\mathcal{L}'	$\mathcal{L}'/n \log_2(\ell)$
San Francisco	7	358081	29067.8	0.0289157	358081	41.2534	0.0000410375
Santa Cruz	6	121733	10605.7	0.0337036	121733	31.9484	0.000101528

FIGURE 8. Non-uniform prior with multiple candidates. Data from the San Francisco and Santa Cruz November 2004 Statement of Vote.

5.2. Discussion. Adding in our non-uniform prior significantly decreases the privacy loss according to the data. Still, we caution that even a 5% value for \mathcal{L}'/n is a concern. Does this mean that 1 in 20 voters has their vote revealed, or does it mean that each voter loses a little bit of privacy of their vote? As we have seen, the answer is somewhere in between for the case of precinct tallies: a few voters lose all privacy, while most lose a varying amount of privacy.

In addition, it may at first seem strange that there is any privacy loss at all in our county-only figures. After all, we used the tallies from the county results to form our prior, so why does this not “reveal everything” to the adversary? The answer is that the prior still has some uncertainty as to the exact tallies of the election, so seeing the actual county tallies removes this uncertainty.

6. EXTENDING THE MODEL

So far we have considered a reporting policy in the context of an adversary with only limited information available. In fact, much more information can be had for analyzing voter privacy. For example, voter party registration is available as part of a county’s voter file. For another example, a voter-verified paper audit trail (VVPAT) may reveal partial information about the order of votes or about the language of the ballot. Coney et al. analyze a case where a VVPAT reveals partial information about the order of votes cast [6]. We could also consider the case in which the VVPAT reveals the language in which the ballot was cast, but we leave this for future work.

7. CONCLUSIONS

We showed that releasing precinct tallies can, in some rare but real cases, completely compromise the privacy of voters. We used the framework of Coney et al. as a starting point to define a quantitative privacy framework for measuring the privacy loss from precinct tallies. We then used real election data to quantify the effect precinct tallies have on voter privacy.

There are several possible responses to privacy loss from precinct reports. One response is to do nothing, judging that the privacy loss is small enough or rare enough as to be outweighed by the benefits in election transparency and in compliance with current laws. Another response, as noted by Hall, is to aggregate precinct reports after the election in cases of a unanimous outcome for a precinct [9]. A different response might aggregate precincts before the election to ensure a minimum precinct size is met, reducing the chance of a unanimous outcome. The full spectrum of responses and trade-offs for each response is beyond the scope of this work. To evaluate these trade-offs and make choices between responses, however, we need a framework for measuring privacy loss.

Another issue is the possibility of matching voters to a particular machine used, as the election audit process may reveal per-machine tallies. The Riverside County registrar explicitly raised per-machine tallies as another part of the concern over releasing precinct tallies [4]. Just as in the whole-precinct case, privacy issues in per-machine tallies could be addressed in several ways, and we need a framework for measuring privacy loss to choose between them.

Therefore, some framework must be used to evaluate privacy loss. How does our approach stack up? With respect to our results, several questions present themselves.

First, it is not clear how to interpret x bits of privacy loss. Our framework uses Shannon entropy to quantify privacy loss, which is an average notion and may or may not say anything about the privacy of an individual voter. As an illustration, how do we compare a reporting policy that completely reveals one person's vote to a reporting policy that reveals a small amount about everyone's vote? With our approach, it is possible for these two policies to have the same value of \mathcal{L}' . We might address this by using different notions of entropy, for example min-entropy, guessing entropy, or Renyi entropy. Wagner gives an overview of these different types of entropies, referring to Cachin's thesis [11, 2].

Second, do different values of \mathcal{L}' impose a total ordering on voting systems and reporting policies? That is, will a lower value of \mathcal{L}' always be "more private" than a higher value? If the answer is yes, then we can fairly compare different proposals for releasing voting information by comparing \mathcal{L}' . This is not a new problem, of course, as this question is fundamental to the entire idea of a voting performance rating. Still, the answer is of interest for our value \mathcal{L}' .

Finally, we note that in the Coney et al. framework, and in our generalization, no useful voting system can be perfectly private - i.e. no voting system can have a privacy loss of zero [6]. Consider a voting system that reveals only the winner of the election and no other information. In the case of an odd number of voters and two candidates, this system has 1 bit of privacy loss. In the case of an even number of voters, the system leaks at most 2 bits and the loss goes to 1 as the number of voters increases. Any useful voting system must report at least this much information, but is there a way to build a system that reveals at most this much?

These questions and others like them are key to formulating quantitative analysis of the tension between transparency and privacy in public elections. At this time, we do not have the answers, and we are not sure we even have all the right questions. Therefore, we cannot recommend making policy decisions based on our current calculations. What we need, instead, is more discussion and exploration of what it means to measure privacy loss and how privacy loss can be measured in a principled way.

8. ACKNOWLEDGMENTS

We thank Poorvi Vora and David Wagner for suggesting the problem of voter privacy and discussions about the privacy loss framework. Joseph Lorenzo Hall provided valuable feedback on election procedures and early drafts of this work. We thank Naveen Sastry for helpful discussions. We thank the San Francisco Registrar, Margaret Morrison from the Santa Cruz Registrar, and Alex Samberg for discussions about vote procedures and the voter file. Ilya Mironov pointed us to the related work by Dwork et al., and Shuchi Chawla graciously shared her slides from a talk about her work. We thank Nick Hopper for reviewing a draft on short notice. Finally, we thank the anonymous reviewers for their helpful comments.

REFERENCES

- [1] Wikipedia. Visited 5 May 2005. Duverger's law, 2005. http://en.wikipedia.org/wiki/Duverger's_Law Cites Maurice Duverger, "Factors in a Two-Party and Multiparty System," in *Party Politics and Pressure Groups* (New York: Thomas Y. Crowell, 1972), pp. 23-32.
- [2] C. Cachin. Entropy measures and unconditional security in cryptography, 1997. PhD thesis, ETH Zurich, <ftp://ftp.inf.ethz.ch/pub/publications/dissertations/th12187.ps.gz>.
- [3] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Towards privacy in public databases. In *Theory of Cryptography Conference*, 2005.

- [4] NCTimes.net Chris Bagley, Staff Writer. Results won't be posted at polls, 2006. <http://www.msnbc.msn.com/id/12863950/from/RL.4/>.
- [5] Federal Elections Commission. Federal elections 2000: 2000 presidential popular vote summary table, December 2001. <http://www.fec.gov/pubrec/fe2000/prespop.htm>.
- [6] Lillie Coney, Joseph L. Hall, Poorvi L. Vora, and David Wagner. Towards a privacy measurement criterion for voting system. In *National Conference on Digital Government Research*, 2005.
- [7] State of California County of San Francisco. November 2004 statement of vote, 2004. <http://web.sfgov.org/site/uploadedfiles/election/Guides/SOV041102.xls>.
- [8] State of California County of Santa Cruz. November 2004 statement of vote, 2004. <http://www.votescount.com/nov04/sov.exe>.
- [9] Joseph Lorenzo Hall. Complications with sequoia's VVPAT, tapes in california counties, 2006. <http://josephhall.org/nqb2/index.php/2006/05/31/casequoiaavvpat>.
- [10] W. L. Pelote, L. Berger, B. Cavala, B. Givens, J. Hayes, V. Salazar, and D. Wong. Task force on voter privacy final report. california office of the secretary of state., 2004. http://www.ss.ca.gov/elections/voter_privacy_final_report/intro_tfvf_final_report.pdf.
- [11] D. Wagner. Re: Entropy vs work (mistake in crypto FAQ?), 1998. Post to sci.crypt newsgroup. <http://www.cs.berkeley.edu/~daw/my-posts/entropy-measures>.

APPENDIX A. APPROXIMATING \mathcal{L}'

In Section 3, we defined

$$\mathcal{L}' = n - \sum_{i=1}^m \log_2 \binom{n_i}{k_i}.$$

In general, the above formula $\mathcal{L}' = n - \log_2 \binom{n_1}{k_1} - \dots - \log_2 \binom{n_m}{k_m}$ is difficult to compute for realistic values of n_i and k_i , since precincts may contain hundreds of voters and k_i is often approximately $n_i/2$, hence $\binom{n_i}{k_i}$ may be extremely large. However, in these cases, we may use Stirling's approximation for $\binom{n}{k}$, which is fairly accurate. We approximate

$$\begin{aligned} \binom{n}{k} &= \frac{n!}{k!(n-k)!} \\ &\approx \frac{n^n e^{-n} \sqrt{2\pi n}}{\left(k^k e^{-k} \sqrt{2\pi k}\right) \left((n-k)^{n-k} e^{k-n} \sqrt{2\pi(n-k)}\right)} \\ &= \frac{n^n}{k^k (n-k)^{n-k}} \sqrt{\frac{n}{2\pi k(n-k)}} \end{aligned}$$

This simplifies our computations significantly, since we can make use the logarithmic identities to yield

$$\begin{aligned} \mathcal{L}' &\approx n - \sum_{i=1}^m \log_2 \left(\frac{n_i^{n_i}}{k_i^{k_i} (n_i - k_i)^{n_i - k_i}} \sqrt{\frac{n}{2\pi k(n-k)}} \right) \\ &= n - \sum_{i=1}^m \left(n_i \log_2(n_i) - k_i \log_2(k_i) - (n_i - k_i) \log_2(n_i - k_i) + \frac{\log_2 n_i - \log_2 2\pi k_i (n_i - k_i)}{2} \right) \end{aligned}$$

Note that, in general, the last term will be negligible, hence our approximation becomes

$$\mathcal{L}' \approx n - \sum_{i=1}^m (n_i \log_2(n_i) - k_i \log_2(k_i) - (n_i - k_i) \log_2(n_i - k_i)).$$

APPENDIX B. APPROXIMATING \mathcal{L}' FOR MULTIPLE CANDIDATES

We may, of course, use the same approximation techniques as in the previous section. However, we must take care since for many third parties, $k_{i,j}$ will be fairly small, which makes the approximation error increase. Approximating $\log_2(n!) \approx n \log_2(n) - \log_2(e)n$, we get

$$\begin{aligned} \mathcal{L}' &= n \log_2 \ell - \sum_{i=1}^m \left(\log_2(n_i!) - \sum_{j=1}^{\ell} \log_2(k_{i,j}!) \right) \\ &\approx n \log_2 \ell - \sum_{i=1}^m \left((n_i \log_2(n_i) - \log_2(e)n_i) - \sum_{j=1}^{\ell} (k_{i,j} \log_2(k_{i,j}) - \log_2(e)k_{i,j}) \right) \\ &= n \log_2 \ell - \sum_{i=1}^m \left(n_i \log_2(n_i) - \sum_{j=1}^{\ell} k_{i,j} \log_2(k_{i,j}) \right) \end{aligned}$$

DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE, UNIVERSITY OF CALIFORNIA, BERKELEY;
BERKELEY, CALIFORNIA 94720