# PRIVACY-PRESERVING ELECTRONIC VOTING

Justin ZHAN, Stan MATWIN, and Li-Wu CHANG

**Abstract:** Privacy is an important issue in electronic voting. Several electronic voting schemes have been developed in the past and some of them provided methods for dealing with privacy protection in the electronic voting system. To further enhance the privacy level, in this paper, we propose a new approach to tackle the privacy problem inherent in the electronic voting system. A privacy measure is proposed and extensive privacy analysis is conducted for the proposed scheme. It is shown via experiments that the proposed method is effective in electronic voting systems.

**Keywords:** Privacy, Security, Randomization, Electronic Voting.

## Introduction

Election is a very important process that allows citizens to choose their government representatives. Paper-form elections have been traditionally used. People normally go to certain places, show their valid identification cards to obtain ballots, then write down their preferences and finally cast their ballots into voting boxes. In this process, election officers have to know the real identification of each person who participates in the vote. To prevent a person from voting multiple times, the officers record all persons' IDs. When new people come to vote, the officers have to check whether they have already voted. This process is not only very inefficient but it also violates the voter's privacy.

The rapid developments in computer and network technologies alter the whole election process. Electronic voting has been developed, where computers are used in the election process enabling greater effectiveness. Especially, on-line voting is the trend in voting system development. Voters do not need to go to a central place to cast their ballots; instead they can stay at home or at any other place with Internet connection to cast their ballots. Obviously, electronic voting or on-line voting[1] is desirable since it is convenient for voters and speeds up the whole election process. However, privacy and security requirements may hinder this desirable election approach from being implemented.

A list of privacy and security requirements were proposed by Neumann.[2] In this paper, we mainly deal with privacy requirements. As pointed out by Neumann,[3] voting privacy means that neither election authorities nor anyone else can link any ballot to the voter who has cast it, and all ballots remain secret while the voting is not completed. Voter anonymity can be achieved by hiding the identity of each voter from her ballot and reverse engineering cannot be made. Although various anonymous schemes have been proposed in the past, no completely satisfactory scheme exists to guarantee voter's anonymity in current electronic voting system. On the other hand, the disclosure of ballots after voting is finished also impairs voter's privacy. Therefore, a stronger privacy requirement not only guarantees the voter's privacy before voting is completed but also after its completion.

Aiming at enhancing voter's privacy, we propose a new scheme. The basic idea of our scheme is to somehow mask the vote before each voter casts her ballot, so that even if an election officer can match the identity of each voter with the actual vote, and divulge the votes after voting is finished, the election officer will still not exactly know the real vote of each voter.

The remainder of this paper is organized as follows. First related work is presented. Next, the authors present their privacy enhancing approach. Then, in the section that follows, the privacy achieved by the proposed method is measured. Experimental results are presented in a further section. This is followed by a discussion of other randomization schemes. Finally, conclusions and future research directions are outlined.

## Related Work

In the early work on electronic voting, Chaum proposed the first multi-party secure election protocol, where a technique based on public key cryptography was presented.[4] However, it cannot prevent someone able to break RSA from tracing ballots back to particular voters. In another publication, Chaum proposed a protocol where a voter's privacy can be ensured if all other voters do not cooperate;[5] voters can guarantee that their ballots can be counted; and voters wishing to disrupt an election can cause only a limited delay before being disenfranchised, unless RSA is broken. Iversen[6] presented a cryptographic scheme that fully conforms to the requirements of holding large scale general elections. By ensuring independence between the voters in that they do not have to be present at the same time or go through several phases together, the scheme preserves the privacy of the voters against any subset of dishonest voters and against any proper subset of dishonest candidates, including the government. Robustness is ensured in that no subset of voters can corrupt or disrupt the election. Benaloh and Tuinstra proposed the first verifiable secrete-ballot election protocols in which participants are unable to prove to others how they voted.[7] Sako and Kilian[8] then proposed a receipt-free voting scheme based on mix-type anonymous channel,[9] with an assumption that there exists a private channel through which the center can send the voter a message without fear of eavesdropping.

Recently, Nguyen, Naini and Kurosawa proposed a formal model for security of verifiable shuffles.[10] The model is general and can be extended to mix-nets and verifiable shuffle decryption. A new efficient verifiable shuffle system based on Paillier encryption scheme was developed and its security was proved. Acquisti presented a voting protocol that protects voters' privacy and achieves universal verifiability, receipt-freeness, and uncoercibility without ad hoc physical assumptions or procedural constraints.[11] The proposed scheme allows voters to combine voting credentials with their chosen votes applying the homomorphic properties of certain probabilistic cryptosystems. A cryptographic randomized response technique[12] is developed by Ambainis, Jakobsson, and Lipmaa[13] to guarantee unconditional privacy for respondents to polls.

In this paper, we propose a flexible randomization scheme for multiple candidate elections.

## Privacy Enhancing Approach

Since its introduction, electronic voting has received a great deal of attention. It is believed to be the major voting method in electronic government. Briefly, electronic voting is the process where voters submit their electronic ballots at a certain location or via Internet; the ballots are transmitted to a back-server where they are collected. After obtaining all valid ballots or the voting day has passed, the back-server counts the number of votes for each candidate. Finally, the candidate who receives certain sufficient amount of votes wins the election.

### *Problem*

We consider the case where there are $n$ political parties participating in the election campaign and there are totally $N$ voters who will cast their vote. Without loss of generality, let us assume that there is only one candidate from a particular party. In other words, there are $n$ candidates and $N$ voters. Since voters are concerned about their ballot's privacy, they do not want to reveal their real votes to anyone including the back server. It is desirable to use some technique to mask the real vote of each voter, but we can still compute the accurate number of counts for each candidate. Based on the above requirement, we propose an estimation scheme. The basic idea of the proposed approach is that each voter randomizes the vote before sending it to the back-server. After the back-server receives the randomized votes, the number of total votes can be estimated with sufficient accuracy.

### *Two-Candidate Randomization Scheme*

In this scheme, let us assume that there is an even number of candidates in the campaign. We randomly separate all the candidates into $n/2$ groups where each group is composed of two candidates. For example, suppose there are four candidates: $C_1$, $C_2$, $C_3$, and $C_4$. We then randomly partition them into two groups, e.g., $C_1$ and $C_4$ in one group; and $C_2$ and $C_3$ in the other group. When voters cast the ballots, they will keep their original votes with a certain probability $\theta$; they will alter their original vote to the other candidate in the same group with probability of $1-\theta$. For example, if Alice wants to vote for $C_1$, she generates a random number; if the number is not greater than a certain value $\theta$, she sends vote for $C_1$ to the back-server; if the number is greater than $\theta$, she then sends vote for $C_4$ to the back-server. If Alice wants to vote for $C_2$, she generates a random number again, if the number is not greater than $\theta$, vote for $C_2$ will be sent to the back-server; otherwise vote for $C_3$ will be sent to the back-server.

Let us assume that candidates $C_i$ and $C_j$ belong to the same group. For convenience, we use the following notation:

- Let Pr(i) denote the real proportion of votes for candidate *i*;
- Let Pr(j) be the real proportion of votes for candidate *j*;
- Let $\Pr^+(i)$ be the proportion of votes for candidate *i* in terms of randomized votes;
- Let $\Pr^+(j)$ be the proportion of votes for candidate *j* in terms of randomized votes.

$\Pr^+(i)$ is contributed by $\Pr(i)$ with a probability $\theta$ and by $\Pr(j)$ with a probability $1-\theta$. $\Pr^+(j)$ is contributed by $\Pr(j)$ with a probability $\theta$ and by $\Pr(i)$ with a probability $1-\theta$.

We can obtain then the following estimation model:

$$\begin{cases} \Pr^+(i) = \Pr(i) * \theta + \Pr(j) * (1-\theta) \\ \Pr^+(j) = \Pr(j) * \theta + \Pr(i) * (1-\theta) \end{cases} \tag{1}$$

What we want to compute from this estimation model is $\Pr(i)$ and $\Pr(j)$. We know that $\theta$, $\Pr^+(i)$ and $\Pr^+(j)$ can be calculated from the randomized votes. Solving the above equations, we can obtain $\Pr(i)$ and $\Pr(j)$. Once we get $\Pr(i)$ and $\Pr(j)$, the number of votes for $C_i$ and $C_j$, denoted by $Vote\_count(C_i)$ and $Vote\_count(C_j)$ respectively, can be computed as follows:

$$\begin{cases} Vote\_Count(C_i) = \Pr(i) * N \\ Vote\_Count(C_j) = \Pr(j) * N, \end{cases}$$

where $N$ is the total number of voters.

To get $\Pr(i)$ and $\Pr(j)$, we have to apply the estimation model presented by Equation 1. How close are the estimated probabilities $\Pr(i)$ and $\Pr(j)$ to the original ones is critical for the election process. In a consequent section, a set of experiments will be conducted to test the proposed scheme.

## Measuring Privacy

In a two-candidate scheme, even if an election officer somehow knows the vote (e.g., that it is for candidate $C_i$) of a particular voter (e.g., Alice), he/she is not sure that the true vote of Alice is for candidate $C_i$ and only knows that Alice votes for $C_i$ with probability of $\theta$. In this section, we develop a privacy measure for the proposed two-candidate scheme. To preserve a fair treatment of all groups, the same $\theta$ values are used for all groups, and the privacy measure will be the same for different groups.

For a vote, the original value can be for candidate *i*, $(C_i)$, or candidate *j*, $(C_j)$; the randomized vote can be for candidate *i*, $(C_i)$, or for candidate *j*, $(C_j)$, as well. The privacy comes from the uncertainty about each voter's original vote given a randomized vote. There are four possible randomization results:

- Original vote is for $C_i$, but the vote after randomization is for $C_i$;
- Original vote is for $C_i$ but the vote after randomization is for $C_j$;
- Original vote is for $C_j$ but the vote after randomization is for $C_i$; and
- Original vote is for $C_j$, however, the vote after randomization is for $C_j$.

Let us adopt the following notation:

- Let $X_m$ be the original vote;
- Let $Y_m$ stand for the vote after randomization;
- Let $W_m$ be the probability that the original vote is $C_i$, that is $\Pr(X_m = C_i)$. The probability that the original vote is $C_j$ will be $(1 - W_m)$, that is $\Pr(X_m = C_j) = 1 - W_m$.

The privacy measure for a two-candidate scheme denoted by $P\_two$ can be derived as follows:

$$P\_two = \Pr(X_m = C_i) * \Pr(Y_m = C_i \mid X_m = C_i) * \Pr(X_m = C_j \mid Y_m = C_i)$$

$$+$$

$$\Pr(X_m = C_i) * \Pr(Y_m = C_j \mid X_m = C_i) * \Pr(X_m = C_j \mid Y_m = C_j)$$

$$+$$

$$\Pr(X_m = C_j) * \Pr(Y_m = C_i \mid X_m = C_j) * \Pr(X_m = C_i \mid Y_m = C_i)$$

$$+$$

$$\Pr(X_m = C_j) * \Pr(Y_m = C_j \mid X_m = C_j) * \Pr(X_m = C_i \mid Y_m = C_j)$$

$$= Component_1 + Component_2 + Component_3 + Component_4$$

The first component contains three parts:

- $\Pr(X_m = C_i)$ is the real probability that a voter votes for $C_i$, which is $W_m$.
- $\Pr(Y_m = C_i \mid X_m = C_i)$ is the probability that a randomized vote is for $C_i$ given the original vote is for $C_i$, which is $\theta$.
- $\Pr(X_m = C_j \mid Y_m = C_i)$ is the probability that an original vote is for $C_j$ given that the randomized vote is for $C_i$. Applying Bayes' rule, we obtain $\Pr(Y_m = C_i \mid X_m = C_j) * \Pr(X_m = C_j) / \Pr(Y_m = C_i)$. $\Pr(X_m = C_j)$ is the probability that the original vote is for $C_j$, which is $1 - W_m$. $\Pr(Y_m = C_i \mid X_m = C_j)$ is the probability that a randomized vote is for $C_i$ given that the original vote is for $C_j$, which is $1-\theta$. As for $\Pr(Y_m = C_i)$, we can expand this term and details for that are shown below:

$$Component_1 = W_m * \theta * \frac{\Pr(Y_m = C_i \mid X_m = C_j) * \Pr(X_m = C_j)}{\Pr(Y_m = C_i)}$$

$$= \frac{\theta * (1-\theta) * W_m * (1 - W_m)}{\Pr(Y_m = C_i \mid X_m = C_i) * \Pr(X_m = C_i) + \Pr(Y_m = C_i \mid X_m = C_j) * \Pr(X_m = C_j)}$$

$$= \frac{\theta * (1-\theta) * W_m * (1 - W_m)}{\theta * W_m + (1-\theta) * (1 - W_m)}$$

Similarly, the other components can be computed as follows:

$$Component_2 = \frac{\theta * (1-\theta) * W_m * (1-W_m)}{\theta * (1-W_m) + (1-\theta) * W_m}$$

$$Component_3 = \frac{\theta * (1-\theta) * W_m * (1-W_m)}{\theta * W_m + (1-\theta) * (1-W_m)}$$

$$Component_4 = \frac{\theta * (1-\theta) * W_m * (1-W_m)}{(1-\theta) * W_m + \theta * (1-W_m)}$$

We then obtain:

$$\begin{aligned} P\_two &= \frac{\theta * (1-\theta) * W_m * (1-W_m)}{\theta * W_m + (1-\theta) * (1-W_m)} + \frac{\theta * (1-\theta) * W_m * (1-W_m)}{\theta * (1-W_m) + (1-\theta) * W_m} \\ &+ \frac{\theta * (1-\theta) * W_m * (1-W_m)}{\theta * W_m + (1-\theta) * (1-W_m)} + \frac{\theta * (1-\theta) * W_m * (1-W_m)}{(1-\theta) * W_m + \theta * (1-W_m)} \qquad (2) \\ &= \frac{2\theta * (1-\theta) * W_m * (1-W_m)}{\theta * W_m + (1-\theta) * (1-W_m)} + \frac{2\theta * (1-\theta) * W_m * (1-W_m)}{(1-\theta) * W_m + \theta * (1-W_m)} \end{aligned}$$

From Equation (2) can be seen that $P\_two$ is determined by two parameters: a control parameter $\theta$ and the original vote distribution $W_m$. What else we can observe from Equation (2) is that $P\_two$ is symmetric with respect to $\theta = 0.5$ and $W_m = 0.5$. For instance, for a given $\theta$, $P\_two$ when $W_m = 0$, 0.1, 0.2, 0.3, and 0.4 is the same as the privacy when $W_m = 1$, 0.9, 0.8, 0.7, and 0.6; for a given $W_m$, $P\_two$ when $\theta =$ 0, 0.1, 0.2, 0.3, and 0.4 is the same as the privacy when $\theta = 1$, 0.9, 0.8, 0.7, and 0.6.

To get a better idea of our privacy measure, we conducted a set of experiments on the original data with various distributions. Specifically, we conducted experiments when $W_m = 0.1$, 0.2, 0.3, 0.4, and 0.5. For each data distribution, we compute privacy value for the cases when $\theta = 0$, 0.1, 0.2, 0.3, 0.4, and 0.5.

As we can see from the results (see Figure 1):

- For a given $W_m$, when $\theta = 0$, the original votes are all changed to the other value in the same group. The original votes are entirely disclosed since an adversary can change all the randomized votes to the original ones. Privacy value is 0; when $\theta$ is away from 0 and approaches 0.5, the randomization probability increases. The level of privacy enhances.

- For a given $\theta$, the privacy level increases with the distribution of the original vote approaching 0.5. The privacy level is at its highest point when $W_m = 0.5$.

We see how privacy changes with varying $\theta$ and $W_m$. In practice, an important issue is how to select a proper value for $\theta$. It cannot only be determined by the privacy level. Accuracy of the results is another critical factor for choosing $\theta$. In the next section, we will conduct a set of experiments and show the relationship between the accuracy and $\theta$.

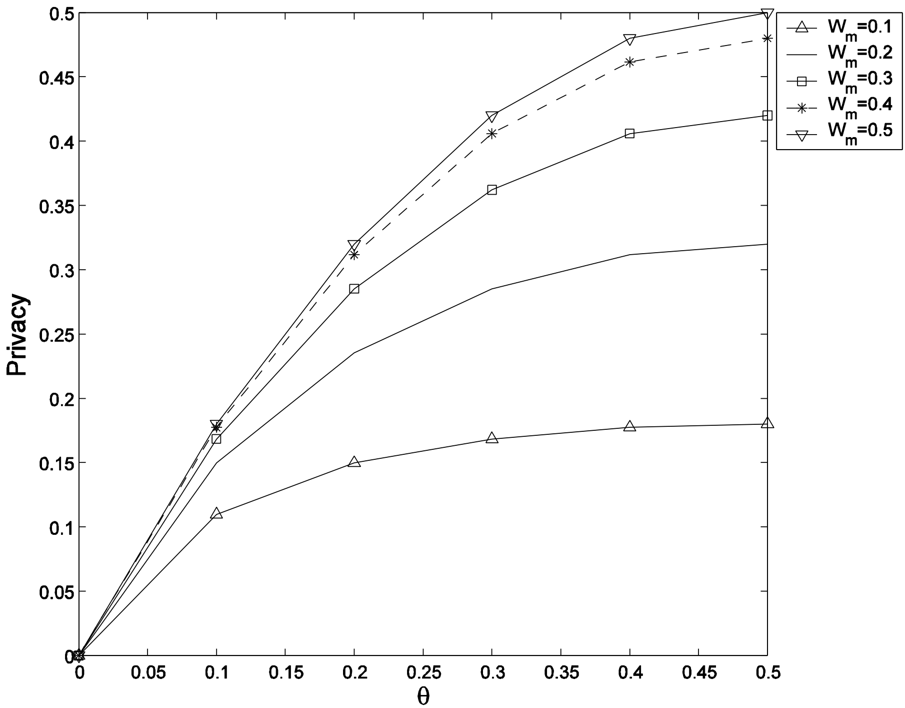


Figure 1: Results from Experiments on the Original Data Set.

## Experimental Results

To evaluate the effectiveness of the proposed scheme, we have conducted a number of experiments on data sets with various distributions. The data sets are randomly generated according to different distributions. In the two-candidate scheme, there are only two candidates $C_i$ and $C_j$ in one group. What we want to know is how close is the estimated proportion of votes to the true proportion of votes for each candidate.

***Experimental Steps***

*1. Data Generation*

The data sets used in the experiments are randomly generated according to the given distribution. The authors evaluate the proposed scheme on nine data sets whose distributions are as follows:

| *Data Sets* | $\Pr(C_i)$ | $\Pr(C_j)$ |
|---|---|---|
| $D_1$ | 0.1 | 0.9 |
| $D_2$ | 0.2 | 0.8 |
| $D_3$ | 0.3 | 0.7 |
| $D_4$ | 0.4 | 0.6 |
| $D_5$ | 0.5 | 0.5 |
| $D_6$ | 0.6 | 0.4 |
| $D_7$ | 0.7 | 0.3 |
| $D_8$ | 0.8 | 0.2 |
| $D_9$ | 0.9 | 0.1 |

*2. $\theta$ Selection*

For $\theta = 0$, 0.1, 0.2, 0.3, 0.4, 0.51, 0.6, 0.7, 0.8, 0.9, and 1, the following steps are performed on data set $D_1$:

- *Step I. Randomization*. For each value in $D_1$, we generate a random number $r$ $(0 \le r \le 1)$ according to an uniform distribution. If $r \le \theta$, the value remains the same; otherwise, the value of *r* will be changed to its opposite. For example, assume that the original value is $C_i(C_j)$, if $r \le \theta$, the value after randomization will still be $C_i(C_j)$; otherwise, the value after randomization will be changed to $C_j(C_i)$. We perform this randomization step for all the values in $D_1$. The data set after the randomization process is denoted by $G_1$.
- *Step II. Estimation*. The model shown by Equation 1 is estimated on the randomized data set $G_1$. We then obtain $\Pr(C_i)$ and $\Pr(C_j)$. Due to the fact

that $\Pr(C_i) = 1 - \Pr(C_j)$, we only record $\Pr(C_i)$ in the resultant tables.

- *Step III. Repeating.* Steps I and II are repeated 50 times, and 50 $\Pr(C_i)s$ are obtained.
- *Step IV. Mean, Variance and Error*. The mean, variance and error percentage of these 50 $\Pr(C_i)s$ are computed.

*3. Compute Mean, Variance and Error for Other Data Sets*

Step 2 is performed on $D_2$, $D_3$, …, and $D_9$.

***Results Analysis***

Tables 1, 2, ..., and 9 show the experimental results on data sets $D_1$, $D_2$, $D_3$, …, and $D_9$, respectively. First, the results are presented followed by a detailed analysis.

Table 1: Results on Data Set $D_1$.

| $\theta$ | *0* | *0.1* | *0.2* | *0.3* | *0.4* | *0.51* | *0.6* | *0.7* | *0.8* | *0.9* | *1.0* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Mean* | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.0995 | 0.0998 | 0.1 | 0.1 | 0.1 | 0.1 |
| *Var* $(*10^{-3})$ | 0 | 0 | 0.0001 | 0.0004 | 0.0016 | 0.1489 | 0.0015 | 0.0004 | 0.0001 | 0 | 0 |
| *Error (%)* | 0 | 0 | 0 | 0 | 0 | 0.05 | 0.02 | 0 | 0 | 0 | 0 |

Table 2: Results on Data Set $D_2$.

| $\theta$ | *0* | *0.1* | *0.2* | *0.3* | *0.4* | *0.51* | *0.6* | *0.7* | *0.8* | *0.9* | *1.0* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Mean* | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.1996 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| *Var* $(*10^{-3})$ | 0 | 0 | 0.0001 | 0.0003 | 0.0012 | 0.1264 | 0.0013 | 0.0002 | 0.0001 | 0 | 0 |
| *Error (%)* | 0 | 0 | 0 | 0 | 0 | 0.04 | 0 | 0 | 0 | 0 | 0 |

Table 3: Results on Data Set $D_3$.

| $\theta$ | *0* | *0.1* | *0.2* | *0.3* | *0.4* | *0.51* | *0.6* | *0.7* | *0.8* | *0.9* | *1.0* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Mean* | 0.3 | 0.3 | 0.3 | 0.3001 | 0.3 | 0.2995 | 0.2999 | 0.3001 | 0.3 | 0.3 | 0.3 |
| *Var* $(*10^{-3})$ | 0 | 0 | 0.0001 | 0.0002 | 0.0012 | 0.1244 | 0.0012 | 0.0002 | 0.0001 | 0 | 0 |
| *Error (%)* | 0 | 0 | 0 | 0.01 | 0 | 0.05 | 0.01 | 0.01 | 0 | 0 | 0 |

Table 4: Results on Data Set $D_4$ .

| $\theta$ | *0* | *0.1* | *0.2* | *0.3* | *0.4* | *0.51* | *0.6* | *0.7* | *0.8* | *0.9* | *1.0* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Mean* | 0.4 | 0.4 | 0.4 | 0.4001 | 0.4001 | 0.3988 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 |
| *Var* $(*10^{-3})$ | 0 | 0 | 0.0001 | 0.0002 | 0.0010 | 0.1498 | 0.0010 | 0.0002 | 0.0001 | 0.0001 | 0 |
| *Error (%)* | 0 | 0 | 0 | 0.01 | 0.01 | 0.12 | 0 | 0 | 0 | 0 | 0 |

Table 5: Results on Data Set $D_5$ .

| $\theta$ | *0* | *0.1* | *0.2* | *0.3* | *0.4* | *0.51* | *0.6* | *0.7* | *0.8* | *0.9* | *1.0* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Mean* | 0.5 | 0.5 | 0.5 | 0.5 | 0.4999 | 0.5010 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| *Var* $(*10^{-3})$ | 0 | 0 | 0.0001 | 0.0002 | 0.0011 | 0.1316 | 0.0013 | 0.0003 | 0.0001 | 0 | 0 |
| *Error (%)* | 0 | 0 | 0 | 0 | 0.01 | 0.1 | 0 | 0 | 0 | 0 | 0 |

Table 6: Results on Data Set $D_6$ .

| $\theta$ | *0* | *0.1* | *0.2* | *0.3* | *0.4* | *0.51* | *0.6* | *0.7* | *0.8* | *0.9* | *1.0* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Mean* | 0.6 | 0.6 | 0.6 | 0.6 | 0.5998 | 0.6020 | 0.6001 | 0.6 | 0.6 | 0.6 | 0.6 |
| *Var* $(*10^{-3})$ | 0 | 0 | 0.0001 | 0.0004 | 0.0021 | 0.2469 | 0.0017 | 0.0003 | 0.0001 | 0 | 0 |
| *Error (%)* | 0 | 0 | 0 | 0 | 0.02 | 0.2 | 0.01 | 0 | 0 | 0 | 0 |

Table 7: Results on Data Set $D_7$ .

| $\theta$ | *0* | *0.1* | *0.2* | *0.3* | *0.4* | *0.51* | *0.6* | *0.7* | *0.8* | *0.9* | *1.0* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Mean* | 0.7 | 0.7 | 0.7 | 0.6999 | 0.6999 | 0.7015 | 0.7 | 0.7001 | 0.7 | 0.7 | 0.7 |
| *Var* $(*10^{-3})$ | 0 | 0.0001 | 0.0001 | 0.0003 | 0.0018 | 0.144 | 0.0016 | 0.0004 | 0.0001 | 0 | 0 |
| *Error (%)* | 0 | 0 | 0 | 0.01 | 0.01 | 0.15 | 0 | 0.01 | 0 | 0 | 0 |

Table 8: Results on Data Set $D_8$ .

| $\theta$ | *0* | *0.1* | *0.2* | *0.3* | *0.4* | *0.51* | *0.6* | *0.7* | *0.8* | *0.9* | *1.0* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Mean* | 0.8 | 0.8 | 0.8 | 0.8 | 0.7999 | 0.8018 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 |
| *Var* $(*10^{-3})$ | 0 | 0 | 0.0001 | 0.0002 | 0.0010 | 0.1498 | 0.0010 | 0.0002 | 0.0001 | 0.0001 | 0 |
| *Error (%)* | 0 | 0 | 0 | 0 | 0.01 | 0.18 | 0 | 0 | 0 | 0 | 0 |

Table 9: Results on Data Set $D_9$ .

| $\theta$ | *0* | *0.1* | *0.2* | *0.3* | *0.4* | *0.51* | *0.6* | *0.7* | *0.8* | *0.9* | *1.0* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Mean* | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9010 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 |
| *Var* $(*10^{-3})$ | 0 | 0 | 0.0001 | 0.0004 | 0.0020 | 0.2500 | 0.0022 | 0.0004 | 0.0001 | 0.0001 | 0 |
| *Error (%)* | 0 | 0 | 0 | 0 | 0 | 0.1 | 0 | 0 | 0 | 0 | 0 |

*Analysis of Mean and Variance*

It can be seen from the tables with the results that when $\theta = 0$ and $\theta = 1$, the estimated proportion of votes is exactly the same as the original proportion, and the variance is 0. This is due to the fact that the randomization process for these two cases does not hide the original votes. When $\theta$ deviates from 1 and 0 and approaches 0.5, the level of randomization increases, and as a result the original information is better disguised. Therefore, the mean of the estimated proportion may deviate from the original proportion and the variance has a trend of becoming larger. Note that when $\theta = 0.5$, the estimation model cannot be applied since Equation (1) does not have a solution in this case. Therefore a value $\theta = 0.51$, instead of 0.5, is used in the experiments.

*Error Rate Analysis*

Estimation error is an important factor in electronic voting. The estimated proportion should not differ very much from the true proportion. Otherwise, the proposed scheme cannot be applied in real electronic voting. Let us use the upcoming (for the time of writing) presidential elections in the United States for illustration. Assume that the true vote proportion for Kerry is 55% and he should be elected as the new president, but the estimated proportion is 45% and he loses the campaign in result. Obviously, it leads to a serious problem. Therefore, the estimated proportion has to be very close to the original proportion. And the most desirable case is when they are the same. As we can see from the results, for most of the $\theta$ values, the estimated proportion is the same as the true proportion. From accuracy point of view, the proposed two-candidate scheme is efficient in these cases.

There are also some limitations in the proposed scheme. Before election, a threshold needs to be agreed among all candidates. Since there are error rates for some cases, we need to find a threshold that is significantly greater than the possible error rates. The error rate reaches its highest point when $\theta = 0.51$ for all data sets. As we can see the highest error rate for $D_1$ is 0.05%, for $D_2$ is 0.04%, for $D_3$ is 0.05%, for $D_4$ is 0.12%, for $D_5$ is 0.1%, for $D_6$ is 0.2%, for $D_7$ is 0.15%, for $D_8$ is 0.18%, and for

$D_9$ is 0.1%. The threshold has to be significantly higher than the highest possible error rate, e.g., 0.2%.

Accuracy and privacy are complementary goals. Given $W_m$, the best privacy is achieved when the control parameter $\theta$ is 0.5; however, the accuracy will be worst in this case. The best accuracy is attained when the control parameter $\theta$ is 0 or 1, however, the privacy is at its lowest level then. Trade-offs are also applied when $\theta$ has a value between 0 and 1. In practice, how to select $\theta$ depends on our primary goal. If we want the results to be very precise, we have to choose values near 1 or 0; in contrast, if privacy is the primary goal, we choose values near 0.5.

## Extension of the Randomization Scheme

### *Three-Candidate Scheme*

In the two-candidate scheme, the randomization process is applied between two candidates since each group contains two candidates. The number of candidates within each group can be increased. In this section, the case when the number of candidates within each group is three is considered. Without loss of generality, let us assume that *n* can be perfectly divided by 3. We randomly separate all the candidates into *n/*3 groups where each group contains three candidates. Suppose that $C_i$, $C_j$ and $C_k$ belong to the same group. When voters cast their ballots, they keep their original votes with a probability $\theta_1$, they alter their original vote to other candidate with probability $\theta_2$ and $\theta_3$, respectively. For instance, if Alice's original vote is for $C_i$, instead of directly sending the vote for $C_i$ to the back-server, she generates a random number $r_1$. If $r_1 \le \theta_1$, she sends a vote for $C_i$ to the back-server. If $\theta_1 \prec r_1 \le \theta_2$, she sends a vote for $C_j$ to the back-server. If $r_1 \succ \theta_2$, she sends a vote for $C_k$ to the back-server. In other words, she keeps her original vote for $C_i$ with a probability $\theta_1$, modifies her original vote to $C_j$ with a probability $\theta_2$ and to $C_k$ with a probability $\theta_3 = (1 - \theta_1 - \theta_2)$.

Let us use the following notation:

- $\Pr(i)$ is the real proportion of votes for candidate $i$ $(C_i)$.
- $\Pr(j)$ is the real proportion of votes for candidate $j$ $(C_j)$.
- $\Pr(k)$ is the real proportion of votes for candidate $k$ $(C_k)$.
- $\Pr^+(i)$ is the proportion of votes for candidate *i* in randomized votes.
- $\Pr^+(j)$ is the proportion of votes for candidate *j* in randomized votes.

- $\Pr^{+}(k)$ is the proportion of votes for candidate *k* in randomized votes.

$\Pr^{+}(i)$ is contributed by $\Pr(i)$ with probability $\theta_1$, $\Pr(j)$ with probability $\theta_3$, and $\Pr(k)$ with probability $\theta_2$. $\Pr^{+}(j)$ is contributed by $\Pr(j)$ with probability $\theta_1$, $\Pr(i)$ with probability $\theta_2$, and $\Pr(k)$ with probability $\theta_3$. $\Pr^{+}(k)$ is contributed by $\Pr(k)$ with probability $\theta_1$, $\Pr(j)$ with probability $\theta_2$, and $\Pr(i)$ with probability $\theta_3$.

The estimation model can be built as follows:

$$\begin{cases} \Pr^{+}(i) = \Pr(i) * \theta_1 + \Pr(j) * \theta_3 + \Pr(k) * \theta_2 \\ \Pr^{+}(j) = \Pr(i) * \theta_2 + \Pr(j) * \theta_1 + \Pr(k) * \theta_3 \\ \Pr^{+}(k) = \Pr(i) * \theta_3 + \Pr(j) * \theta_2 + \Pr(k) * \theta_1 \end{cases} \tag{3}$$

In the above model $\theta_1$, $\theta_2$ and $\theta_3$ are known. $\Pr^{+}(i)$, $\Pr^{+}(j)$ and $\Pr^{+}(k)$ can be computed directly from the randomized votes. We can then solve the above model and obtain $\Pr(i)$, $\Pr(j)$ and $\Pr(k)$. The total number of votes for candidate *i* is $\Pr(i) * N$, for candidate *j* is $\Pr(j) * N$ and for candidate *k* is $\Pr(k) * N$.

***n-Candidate Scheme***

In general, we can treat all the candidates in one group. We call it n-candidate scheme. Voters keep their true votes with a probability $\theta_1$ and alter the vote to the other candidates with probabilities $\theta_2$, $\theta_3$, …, and $\theta_n$, respectively.

Let us assume the following notation:

- $\Pr(C_m)$ $(m = 1, 2, \cdots, n)$: the real proportion of votes for candidate *m* $(C_m)$.
- $\Pr^{+}(C_m)$ $(m = 1, 2, \cdots, n)$: the proportion of votes for candidate *m* in randomized votes.

Then the estimation model will look as follows:

$$\begin{cases} \Pr^{+}(C_1) = \Pr(C_1) * \theta_1 + \Pr(C_2) * \theta_n + \Pr(C_3) * \theta_{n-1} + \Pr(C_4) * \theta_{n-2} + \ldots + \Pr(C_n) * \theta_2 \\ \Pr^{+}(C_2) = \Pr(C_1) * \theta_2 + \Pr(C_2) * \theta_1 + \Pr(C_3) * \theta_n + \Pr(C_4) * \theta_{n-1} + \ldots + \Pr(C_n) * \theta_3 \\ \Pr^{+}(C_3) = \Pr(C_1) * \theta_3 + \Pr(C_2) * \theta_2 + \Pr(C_3) * \theta_1 + \Pr(C_4) * \theta_n + \ldots + \Pr(C_n) * \theta_4 \\ \qquad \qquad \qquad \qquad \ldots\ldots \\ \Pr^{+}(C_n) = \Pr(C_1) * \theta_n + \Pr(C_2) * \theta_{n-1} + \Pr(C_3) * \theta_{n-2} + \Pr(C_4) * \theta_{n-3} + \ldots + \Pr(C_n) * \theta_1 \end{cases}$$

In the above estimation model, we can estimate $\Pr(C_m)$ $(m=1,2,\cdots,n)$ since $\theta_m$ $(m=1,2,\cdots,n)$ are known and $\Pr^+(C_m)$ $(m=1,2,\cdots,n)$ can be directly computed from the collected randomized votes. The total number of votes for candidate *m* is then $\Pr(C_m)*N$ $(m=1,2,\cdots,n)$.

We see that the number of equations in the estimation model of the n-candidate scheme is equal to the number of candidates participating in the campaign. Although the two-candidate scheme is much simpler, other candidate schemes can certainly be used. In practice, different schemes can be combined together. For instance, since the number of candidates may not be perfectly divided by two or three, we can combine two- and three-candidate schemes. One possibility is to separate all the candidates into *g* groups with *g-1* groups containing two candidates and with 1 group containing three candidates.

## Conclusions and Future Work

Electronic voting is an efficient way of performing governmental elections. Especially, the controversial year 2000 elections in the United States of America made people realize the importance of electronic voting. Although electronic voting or online voting systems can significantly improve the efficiency of voting, security and privacy violations may prevent them from being implemented. To preserve privacy in electronic voting, a privacy protection method has been introduced in this paper. In the proposed technique, voter's vote is randomized before sending it to back-server. The performed experiments have illustrated that the election results are still very accurate although the original votes have been hidden. A privacy measure has been developed and a privacy analysis conducted. Trade-offs between privacy and accuracy have been discussed. In the future, the authors intend to combine the proposed technique with other privacy and security protection techniques for electronic voting. The approach will also be extended to other e-government services.

**Notes:**

1 Aviel D. Rubin, "Security Considerations for Remote Electronic Voting over the Internet" (22 June 2001), <http://avirubin.com/e-voting.security.pdf> (15 November 2004).

2 Peter G. Neumann, "Security Criteria for Electronic Voting," (paper presented at the 16th National Computer Security Conference, Baltimore, Maryland, September 1993), 20-23.

3 Neumann, "Security Criteria for Electronic Voting."

4 David L. Chaum, "Untraceable Electronic Mail, Return Address, and Digital Pseudonyms," *Communication of the ACM* 24, no. 2 (1981): 84-88.

5 David L. Chaum, "Elections with Unconditionally-Secret Ballots and Disruption Equivalent to Breaking RSA," in *Advances in Cryptology - EUROCRYPT '88* (Berlin, 25-27 May

1988), *Lecture Notes in Computer Science* 330, ed. Christoph G. Gunther (Springer-Verlag, 1988), 177-182.

[6] Kenneth R. Iversen, “A Cryptographic Scheme for Computerized General Elections,” in *Advances in Cryptology -- CRYPTO '91* (11-15 August 1991), *Lecture Notes in Computer Science* 576, ed. J. Feigenbaum (Berlin: Springer-Verlag, 1992), 405-419.

[7] Josh D. Benaloh and Dwight Tuinstra, “Receipt-Free Secret-Ballot Elections (extended abstract),” in *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing STOC’94* (New York, May 1994), 544-553.

[8] Kazue Sako and Joe Kilian, “Receipt-Free Mix-Type Voting Scheme – a Practical Solution to the Implementation of a Voting Booth,” in *Advances in Cryptology - EUROCRYPT’95* (Saint-Malo, France, 21-25 May 1995), *Lecture Notes in Computer Science* 921, ed. Louis C. Guillou and Jean-Jacques Quisquater (Berlin: Springer-Verlag, 1995), 393-403.

[9] Choonsik Park, Kazutomo Itoh, and Kaoru Kurosawa, “All-Nothing Election Scheme and Anonymous Channel,” in *Advances in Cryptology - EUROCRYPT’93* (Lofthus, May 1993), *Lecture Notes in Computer Science* 765, ed. T.Helleseth (Berlin: Springer-Verlag, 1993), 248-259; Chaum, “Untraceable Electronic Mail, Return Address, and Digital Pseudonyms.”

[10] Lan Nguyen, Reihaneh Safavi-Naini, and Kaoru Kurosawa, “Verifiable Shuffles: A Formal Model and a Paillier-Based Efficient Construction with Provable Security,” in *Applied Cryptography and Network Security: Second International Conference ACNS 2004* (Yellow Mountain, China, 8-11 June 2004), *Lecture Notes in Computer Science* 3089, ed. Markus Jakobsson, Moti Yung, and Jianying Zhou (Heidelberg: Springer-Verlag, 2004), 61-75.

[11] Alessandro Acquisti, “Receipt-Free Homomorphic Elections and Write-in Voter Verified Ballots,” Technical Report CMU-ISRI-04-116 (Carnegie Mellon University, School of Computer Science, April 2004), <citeseer.ist.psu.edu/663984.html> (15 November 2004).

[12] Stanley L. Warner, “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias,” *Journal of American Statistical Association* 60, no. 309 (March 1965): 63-69.

[13] Andris Ambainis, Markus Jakobsson, and Helger Lipmaa, “Cryptographic Randomized Response Techniques,” in *Public Key Cryptography – PKC 2004: 7th International Workshop on Theory and Practice in Public Key Cryptography* (Singapore, 1-4 March 2004), *Lecture Notes in Computer Science* 2947, ed. Feng Bao, Robert H. Deng, and Jianying Zhou (Heidelberg: Springer-Verlag, 2004), 425-438.

**JUSTIN ZHAN** is a part-time professor at the School of Information Technology and Engineering, University of Ottawa, Canada. His research interest is privacy and security issues in data mining. *E-mail:* zhizhan@site.uottawa.ca.

**STAN MATWIN** is a professor at the School of Information Technology and Engineering, University of Ottawa, Canada. His research is in machine learning, data mining, and their applications, as well as in technological aspects of Electronic Commerce. *E-mail:* stan@site.uottawa.ca.

**LI-WU CHANG** is a research scientist at Center for High Assurance Computer Systems of Naval Research Laboratory, USA. *E-mail:* lchang@itd.nrl.navy.mil.