New Insight to Preserve Online Survey Accuracy and Privacy in Big Data Era

Joseph K. Liu - ***Institute for Infocomm Research, Singapore***
Man Ho Au - *The Hong Kong Polytechnic University, Hong Kong*
Xinyi Huang - *Fujian Normal University, China*
Willy Susilo – *University of Wollongong, Australia*
Jianying Zhou - *Institute for Infocomm Research, Singapore*
Yong Yu - *University of Electronic Science and Technology of China, China*

Table of Content

- Introduction
- Contribution
- Related Works
- Our System
- Security Analysis
- Efficiency Analysis
- Other Applications
- Conclusion

Introduction – Online Survey

- An Internet survey technique (e.g. Kwik Survey, My3q or Survey Monkey)
- The questionnaires are created in a program for creating web interviews
- It is considered to be a cheaper way of conducting surveys since it does not require any human resources to conduct surveys or telephone interview
- ***Accuracy*** and ***privacy*** should be carefully addressed

Accuracy of Online Survey

- A survey form may collect the interviewee's personal particulars, such as sex, age, salary range and interest
- The interviewer has no way to verify the authenticity of this information
- A 15 years old boy may say that “she” is a 50 years old woman
- Next time he may pretend he is a retired 80 years old man
- There is no way to verify whether these 2 different surveys are from the same source or not (especially if they come from different IP)
- Digital signature provides an easy and convenient way to authenticate the message sender
- No Privacy for digital signature!

Privacy of Online Survey

- Many users are not willing to reveal their real identities to interviewers
- If it is a compulsory requirement for conducting the survey, they will decline the survey invitation
- It maybe the main reason that many existing online survey systems do not compulsorily require interviewees to input their real identifying information
- Or at least no need to verify their information

Contribution

- We provide a new insight to preserve accuracy and privacy in online survey systems. Our proposed system provides the following desirable features:
  - Authentication: It allows only those authenticated or qualified users to take part into the survey
  - Anonymity: No one knows the identity of the user who has submitted the survey
  - Detection of double submission: No one can submit more than once in a single survey event without being detected
  - Unlinkability: Given two surveys from two different events, no one can tell whether they are from the same user
  - Constant Complexity: The complexity of our system is independent to the total number of users in the system.
    - *It is particularly suitable for any system with large user database in the big data analytic era*

Related Works

- There are many ways to resolve the contradiction between user privacy and data accuracy
  - Ring signature
  - Group signature
  - Attribute-based signature

Attribute Based Signature (ABS)

- Traditional signature, a user is represented by his unique *public key* or *identity*.
- In an ABS, a user is represented by a set of *attributes*:
  - User 1: {male, engineer, company A, French}
  - User 2: {female, clerk, company B, German}
  - User 3: {female, programmer, company A, Indian}
  - User 4: {male, engineer, company C, German}
- There is no specific public key / identity for any user
- For signature, one can sign a message with a predicate satisfied by his/her attributes
  - Signature on *m* for {male, engineer}
  - Verifier can confirm the signer is {male}, {engineer}.
  - However, verifier does not know who is the actual signer (user 1 or user 4).
- Privacy is protected
- Authentication is provided

Attribute Based Signature (ABS)

- Is the privacy provided by ABS too *strong*?
- No one can check whether two signatures are produced by the same user
- In our case, the survey system should be better to check whether or not two attribute-based signatures were produced by the same user (Detection of double submission)

Our System – Basic Idea

- Our system is based on ABS
- However, due to the unlinkability property of an ABS scheme, it is not suitable to be used directly
- We add *linkability* to it
- Any verifier is able to detect whether two signatures are generated by the same user within a single survey
- Yet any user that generates two signatures in two different surveys cannot be linked
- The survey centre can discard any double-submitted survey to maintain the accuracy of the result
- We build our system based on [MPR11]
  - [MPR11] H. K. Maji, M. Prabhakaran, and M. Rosulek. Attribute-Based Signatures. In CT-RSA, volume 6558 of LNCS, pages 376{392. Springer, 2011.

Our System – Entities

- Attribute Authority (AA): It is responsible for setting up the public parameters and issuing user secret keys for various attributes
- User: Any entity who has a user secret key is an user. A user can have different attributes
- Survey Centre (SC): It is an organization to organize a survey. It is responsible to define the required policy of the survey, to collect the survey from users and to verify the result

The Construction - Setup

- The AA defines all system parameters and generates the public key and a master secret key

1. Let $\hat{e} : \mathbb{G}_1 \times \mathbb{G}_2 \rightarrow \mathbb{G}_T$ be a bilinear map (defined in Section 3.1) such that $|\mathbb{G}_1| = |\mathbb{G}_2| = |\mathbb{G}_T| = p$ for some prime p. Let g, G be generators of $\mathbb{G}_1$ and $\mathfrak{g}, \mathfrak{h}, h, h_0, \ldots, h_{t_{max}}, H$ be generators of $\mathbb{G}_2$. The value t_{max} is the maximum width of the monotone span programs as defined in Section 3.2. Let $\mathbb{A} = \mathbb{Z}_p^*$ be the universe of attributes.
2. Assume the DDH problem (defined in Section 3.1) is hard in $\mathbb{G}_1$ and $\mathbb{G}_2$. Let $\mathcal{G} : \{0,1\}^* \rightarrow \mathbb{G}_1$, $\mathcal{H} : \{0,1\}^* \rightarrow \mathbb{Z}_p$ be hash functions that will be modeled as random oracles. The system parameters TPK is ($\mathbb{G}_1$, $\mathbb{G}_2$, $\mathbb{G}_T$, $\hat{e}$, p, g, G, $\mathfrak{g}$, $\mathfrak{h}$, h, h_0, ..., $h_{t_{max}}$, H, $\mathcal{H}$, $\mathcal{G}$).

Then it generates the public and master secret keys as follows:

1. Choose $a_0, a, b, c \in_R \mathbb{Z}_p$ and compute: $C = g^c$, $A_0 = h_0^{a_0}$, $A_j = h_j^a$, $B_j = h_j^b$ for $j = 1, \ldots, t_{max}$.
2. Choose $s, v, w, z \in_R \mathbb{Z}_p$ and compute: $U = G^s$, $V = H^v$, $W = H^w$, $Z = H^z$.
3. Set the public key APK as (C, A_0, $\{A_j, B_j\}_{j=1}^{t_{max}}$, U, V, W, Z) and the master secret key ASK as (a, a_0, b, s, v, w, z). Publish both APK and TPK while keep ASK secret.

The Construction - User Key Generation

- The AA issues user secret key to each user, according to different attributes each user possesses

1. The user with an attribute set $\mathcal{A} \in \mathbb{A}$ randomly selects $L, r_L \in \mathbb{Z}_p$ and computes $C_L = \mathfrak{g}^L \mathfrak{h}^{r_L} \in \mathbb{G}_2$ and sends C_L to the AA.
2. The AA randomly chooses $K_{base} \in_R \mathbb{G}_1$, $r \in_R \mathbb{Z}_p$ and uses the master secret key ASK to compute: $K_0 = K_{base}^{\frac{1}{a_0}}$, $K_u = K_{base}^{\frac{1}{a+bu}}$ $\forall u \in \mathcal{A}$, $R = G^r$, $S = G^{z-rv} K_{base}^{-w}$, $T = (H C_L^{-s})^{\frac{1}{r}}$.
3. The AA returns K_{base}, K_0, $\{K_u\}_{u \in \mathcal{A}}$, R, S, T to the user.
4. The user parses his user secret key $SK_{\mathcal{A}}$ as $(K_{base}, K_0, \{K_u\}_{u \in \mathcal{A}}, R, S, T, L, r_L)$.

The Construction - Survey Submission

- The SC defines a survey event and a policy such that only those users that fulfill the policy with their attributes can participate this survey.
- The user submits the survey data together with the corresponding signature signed with his user secret key through an anonymous channel to the SC

The user executes the following steps with his user secret key $SK_{\mathcal{A}}$:

1. Compute $\mu = \mathcal{H}(m||\Upsilon)$ and $\tau = \mathcal{G}(event)^L$.
2. Pick $r_0 \in_R \mathbb{Z}_p^*$, $r_1, \ldots, r_\ell \in_R \mathbb{Z}_p$ and compute $Y = K_{base}^{r_0}$, $W = K_0^{r_0}$, $S_i = (K_{u(i)}^{v[i]})^{r_0}(Cg^{\mu})^{r_i}(\forall i \in [\ell])$, $P_j = \prod_{i=1}^{\ell}(A_j B_j^{u(i)})^{\mathcal{M}_{ij} \cdot r_i}(\forall j \in [t])$.
3. Compute Π_τ as a non-interactive zero-knowledge proof-of-knowledge of the values $(R, S, T, K_{base}, r_0, L, r_L)$ satisfying the following relation:

$$\begin{aligned} \hat{e}(R,V)\hat{e}(S,H)\hat{e}(K_{base},W) &= \hat{e}(G,Z) \quad \wedge \\ \hat{e}(R,T)\hat{e}(U,\mathfrak{g}^L\mathfrak{h}^{r_L}) &= \hat{e}(G,H) \quad \wedge \\ Y &= K_{base}^{r_0} \quad \wedge \\ \tau &= \mathcal{G}(event)^L. \end{aligned}$$

4. Submit the survey data m with its signature $\sigma = \left(Y, W, \{S_i\}_{i\in[\ell]}, \{P_j\}_{j\in[t]}, \tau, \Pi_\tau\right)$ to the SC.

The Construction – Validity Checking

- Upon received the survey, the SC checks its validity. The checking consists of two parts:

1. Signature Verification:
   (a) Convert the policy Υ such that $\Upsilon(\mathcal{A}) = 1$ to its corresponding monotone span program $\mathcal{M} \in \mathbb{Z}_p^{\ell \times t}$, with row labeling function $u : [\ell] \to \mathbb{A}$.
   (b) Compute $\mu = \mathcal{H}(m||\Upsilon)$ and check if $\hat{e}(W, A_0) \stackrel{?}{=} \hat{e}(Y, h_0)$ and

$$\prod_{i=1}^{\ell} \hat{e}(S_i, (A_j B_j^{u(i)})^{\mathcal{M}_{i,j}}) \stackrel{?}{=} \begin{cases} \hat{e}(Y, h_1)\hat{e}(Cg^{\mu}, P_1), & \text{for } j = 1. \\ \hat{e}(Cg^{\mu}, P_j), & \text{for } j > 1. \end{cases}$$

   (c) Checks if Π_τ is a valid proof. The verification of Π_τ is also shown in Appendix A.

   If all equalities hold and the proof is correct, it outputs ACCEPT and proceeds to the second part. Otherwise it outputs REJECT.
2. Double Submission Checking: The SC extracts τ from σ and checks its database whether any other signatures for this survey *event* also contain the the same τ. If yes, that means the user has double submissions. It then outputs REJECT. Otherewise, it outputs ACCEPT and stores the data and signature into its database.

Security Analysis

- Security against Unforgeability Attack
  - The attacker acts as an unauthorized user (who does not possess the required attributes) who tries to submit a survey to the SC for being accepted
  - Each survey response has to be accompanied with a properly created attribute-based signature and in our system, only authorized users are issued the signing keys
  - The ABS of [MPR11] is unforgeable → Our signature is unforgeable
- Security against Anonymity Attack
  - The attacker acts as the AA colluded with the SC who tries to find out the identity of the user of a particular submission
  - The ABS of [MPR11] is anonymous
  - Π_τ leaks no information due to its zero-knowledgeness
  - Since L is never shown in plain and is protected by the perfect hiding property of the Pedersen commitment, it again leaks no information about the survey participant

Security Analysis

- Security against Linkability Attack
  - The attacker acts as an authorized user who tries to submit more than one survey to the SC for being accepted in a single survey event
  - Each authorized user in our system is given only one “certified signing key" only and thus for each survey, he/she can only generate one unique tag
  - This is due to the fact that the non-interactive zero-knowledge proof is sound
- Unlinkability (for different users) Attack
  - The attacker acts as an authorized user who tries to submit some surveys to link with other surveys submitted by honest users
  - The attack may have intention to do so in order to remove other undesirable results submitted by other users
  - In order to use a tag, the attacker has to produce the zero-knowledge proof
  - The attacker either produces a fake proof or has to know the value of L that is used to generate the tag
  - The former is computationally impossible under the soundness property of the zero-knowledge proof
  - The latter is computationally impossible under the discrete logarithm assumption

Efficiency Analysis - Generic

- We use t_{max} to represent the maximum width of the monotone span program
- $|A|$ to represent the number of attributes a user has
- t and ℓ to represent the width and length of the monotone span program converted from the signing claim policy

	ASetup	AttrGen	Sign	Verify
Group $\mathbb{G}_1$ / $\mathbb{G}_2$ exponentiation (pre-processed)	$6 + 2\,t_{max}$	3	12	0
Group $\mathbb{G}_1$ / $\mathbb{G}_2$ exponentiation (no pre-processed)	0	$2 + \|\mathcal{A}\|$	$2 + 2\,\ell + t\ell$	$6 + t\ell$
Group $\mathbb{G}_T$ exponentiation (pre-processed)	0	0	7	7
Group $\mathbb{G}_T$ exponentiation (no pre-processed)	0	0	0	2
Pairing (1 element is a constant)	0	0	0	$5 + t$
Pairing (both elements are not constant)	0	0	0	$2 + t\ell$

Efficiency Analysis - Example

Concrete Example. Next we analyze the efficiency of our scheme using the simulation result from jPBC [21] for the following devices:

- A desktop equiped with Intel(R) Core(TM)2 Quad CPU Q6600 2.40GHz, 3 GB RAM, Ubuntu 10.04 as the simulation device.

We measured the performance using a 160-bit secret key in elliptic curve cryptosystem (ECC). It is generally believed that a 160-bit secret key in ECC provides stronger security than a 1024-bit key in RSA.

In the example, we assume the following attributes:

- Sex: {Male}, {Female}
- Marriage Status: {Single}, {Married}, {Divoice}
- Office Location: {United States}, {United Kingdom}, {Australia}, {Japan}, {China}
- Year of Bith: $\{\leq 1960\}$, $\{1961 - 1970\}$, $\{1971 - 1980\}$, $\{1981 - 1990\}$, $\{> 1990\}$
- Department: {Sales}, {Finance}, {Logistic}, {Human Resources}

Efficiency Analysis - Example

Now it plans to carry out some surveys based on the following different cases:

1. All staffs who are based in Japan.
2. All Female staffs who are Married.
3. All Male staffs who are based in Australia and working in the Sales department.
4. All Female staffs who are Single, born after 1990 and based in Japan.
5. All Male staffs who are Married, based in United States, born between 1971-1980 are working in the Finance department.
6. All staffs who are based in *either* Australia *or* China and working in the Sales department.
7. All Female staffs who are based in United Kingdom and working in *either* Finance *or* Human Resources department.

Case	Size of *APK*	**Ur. Key Gen.** running time	Size of se. key	**Survey Subm.** running time	Size of signature	**Val. Check** running time
1	880	138.667	220	127.052	480	225.915
2				182.828	520	333.185
3				350.156	560	506.893
4				517.484	600	747.12
5				721.996	640	1053.839
6				294.38	540	399.65
7				443.116	580	606.631

Other Applications – E-voting

- There are many existing e-voting protocols
- When compared to others, using ABS (with linkability) is more efficient
- For example
  - In a university, each student has the attribute set:
    {*sex, college, department*}
  - Alice: {female, college A, computer science}
  - Bob: {male, college B, computer science}
  - Carol: {female, college B, music}
  - Daniel: {male, college A, history}
  - Elaine: {female, college A, music}
  - Frankie: {male, college B, computer science}

Application – E-voting

- Election 1: A student union election for compute science:
  - Alice, Bob, Frankie
    - Each one generates a signature (the message can be the vote), using the attribute {computer science} and posts to a public board anonymously
    - Double voting can be detected
- Election 2: A female student union election for college A:
  - Alice, Elaine
    - Each one generates a signature (the message can be the vote), using the attributes {female}, {college A} and posts to a public board anonymously
    - Double voting can be detected
    - If Alice votes once in Election 1, and votes once in Election 2, Alice will *not* be detected, as Election 1 and Election 2 are two different events
- Advantage:
  - efficient – only need to generate one signature
  - no need setup for each event – the user secret key can be used in different events
  - no need to know who else are eligible voters

Application – Smart Grid

- a form of electricity network utilizing modern digital technology
  - Two-way communication
  - Electricity company provides power supply to consumer
  - Consumer provides usage data to provider
  - Better utilizing electricity supply
  - Save more energy, save more money

Application – Smart Grid

- Privacy is an important concern for consumers
- By using normal ABS, privacy can be preserved and data can be authenticated:
  - E.g. Each user has the attribute set
    {Region, Building name, number of family members}
  - User A may have {East District, Victoria Building, 4}
  - The company may want to collect data from users living in {East District} and have {4} family members
- However, it cannot be detected if user A sends the data twice (to disturb the statistic data)
  - Double sending cannot be detected
- Event-linkable ABS provides a mechanism to detect double sending

Application – Vehicle Ad Hoc Network (VANET)

- Allow wireless communications between vehicles and roadside infrastructures
- Any vehicle may broadcast a piece of information to other vehicles without going through a central server
- Problem of reliability of information:
  - Bob receives a message from another vehicle reporting some traffic jam a few miles away, he has no idea whether the message is true or not
  - At the beginning, he tries to ignore
  - However, later he receives a number of the same traffic jam message (say, n)
  - If n is reasonably large enough, and these n messages are sent by n vehicles, most likely this message is true
  - The question is:
    - All these messages are sent anonymously, due to privacy concern
    - How can Bob know that these messages are sent by n different vehicles (instead of 1 vehicle sends n times?

Application – Vehicle Ad Hoc Network (VANET)

- Chen et al [9] proposed a solution
  - Threshold Anonymous Announcement (TAA)
  - Each vehicle obtains a token from a trusted party
  - Broadcast an anonymous message signed by this token
  - If a vehicle sends the *same message* twice, the receiver will be able to know these two messages are sent by the same vehicle
- The solution seems very good, however…
  - If the signer slightly changes the message, e.g., change from "`The city area is very congested now.`" to "`Now the city area is very congested.`"
  - they appear as two different messages and thus cannot be linked
- This problem can be solved by using our proposed ELABS
  - Event linkable instead of message linkable

Conclusion

- We provided a new insight to preserve accuracy and privacy in online survey systems simultaneously
- The new insight comes from our proposed system
- We add linkability to a normal ABS
- In addition to online survey systems, we further suggested several other applications that can make use of our new system, including e-voting, smart-grid and vehicular ad hoc networks
- We believe our system is particular suitable for handling big data as the complexity remains constant, regardless to the number of users

~~~ END ~~~
~~~