Enhance Biometric Database Privacy: Defining Privacy-Preserving Drawer Size Standard for the Setbase

Benjamin Justus[1], Frédéric Cuppens[1], Nora Cuppens-Boulahia[1], Julien Bringer[2], Hervé Chabanne[2], and Olivier Cipiere[2]

[1] Lab-STICC, Télécom Bretagne, Cesson Sévigné, France
[2] Morpho, Paris, France

Abstract. Shamir proposed the setbase approach as a means of improving security and privacy of the traditional biometric system. In this paper, we propose privacy-preserving drawer size standards for the biometric setbase. The proposal incorporates database privacy metrics such as k-anonymity and l-diversity into the definition of privacy-preserving drawer size standard for the biometric setbase. We also empirically evaluate the system reliability of the prototype setbase for the purpose of studying the trade-off values between the level of privacy protection and the level of system security.

Keywords: database fragmentation, setbase, k-anonymity, l-diversity, biometric privacy.

1 Introduction

The setbase approach was proposed by Adi Shamir [6] (see also [2]) as a means of improving security and privacy of the traditional biometric system. The traditional biometric database uses a one-to-one data linking between biometric data stored in the biometric database and personal information stored in the identity database. The setbase approach depends on the creation of drawers in which the biometric and personal data are stored. The linking between the identity and the corresponding biometric data takes places at the level of linking identical drawer IDs in the databases. Since each drawer contains multiple records, the identification of an identity is blurred which leads to privacy protection of an individual. With the number of drawers sufficiently large, the probability of an identity theft can be made negligible. Furthermore, the lack of cryptographic mechanisms in a setbase can be seen as another privacy virtue because we do not want to place too much trust in a single entity as for instance the owner of the cryptographic keys.

Given the theoretic interest of Shamir's proposal, there have been hitherto no studies in the literature on how practical shamir's scheme is. Specifically, one would like to know at least in theory:

L. Wang and B. Shafiq (Eds.): DBSec 2013, LNCS 7964, pp. 274–281, 2013.

1. Set of metrics capable of measuring the privacy level of the setbase approach.
2. Value ranges of the important setbase system parameters, such as the number of drawers and possible sizes of a drawer.
3. Impact of the setbase parameters on the level of database privacy.

The importance of choosing the appropriate drawer sizes in a setbase can be seen in the context of preserving privacy of individuals in a biometric database. To make things concrete, we explain the concept using the crime investigation example. In such a scenario, the investigator has gathered some generic information such as the sex, the age group, and the domicile region of a suspect. Furthermore, the investigator based on the biometric data (e.g. fingerprints collected at the crime scene) is able to locate the drawer in which a suspect is located. To identify the person, he must retrieve the identity of a suspect. If the drawer size is small, the investigator would have a little trouble in sieving out entries inside the drawer that satisfy his searching criteria. In the extreme case of a traditional biometric database (one-to-one biometric and identity association), the identity of a suspect would be revealed once a successful bio and identity matching is achieved. The increase of the drawer size would make an investigator's search more difficult. Thus strong privacy in a setbase requires a large drawer size. The downside of a large drawer size is that it would expose the system to a higher chance of drawer ID collision attacks (see section 3). The appropriate choices of the drawer size for a setbase require a judicious way of balancing between privacy and security.

1.1 Our Contribution

The starting point of our project is a prototype implementation of the biometric setbase in the context of studying its suitability for the integration with the French national identity cards. There had been many propositions in the French National Assembly advocating the protection of sensitive information such as those contained in the national identity cards [1,5]. The main goal of our project is on the one hand to measure the feasibility of the setbase approach for a real world implementation, and on the other hand to quantify as accurately as possible how much privacy is enhanced by using the setbase approach.

In order to measure the privacy level of the setbase, we had to develop a drawer size standard that depends on a set of well known database privacy metrics. The current paper relies on the concept of k-anonymity and l-diversity as a means of measuring the privacy level of data tables contained in the database. Our study is based on the French population data released from National Institute of Statistics and Economic Studies. The present study shows that the drawer size standard proposed in the paper permits feasible setbase solution for the integration with the traditional biometric database currently in use.

1.2 Paper Outline

The paper is organized as follows. The proposed privacy-preserving drawer size formula is presented in section 2. We evaluate the prototype system performance

empirically in section 3. Section 4 provides a detailed analysis of the trade-off between security and privacy when deciding the drawer size.

1.3 Notation

We use the following notations throughout the paper:

NT **:** the number of drawers that is maintained by the system
TT **:** the size of a drawer that is maintained by the system

2 Database Tracing Modeling

The aim of this section is to provide some reasonings behind the privacy-preserving drawer size formula:

$$TT = Tol(k) \cdot E(X)_{\mathbf{P}} + N_{expire} \tag{1}$$

where $Tol(k)$ is a linear function depending on the k-anonymity parameter (see [4,3] for the definition). N_{expire} is a fixed number that represents the number of expired entries in the drawer. It is a numerical parameter belonging to the drawers due to the definite expiration date of the biometric data. The value N_{expire} is usually determined by the relevant government policies in the public domains. The random variable X represents the number of sequential searches required before one hits an identity record that satisfies the specified sex, age group, and domicile region criteria. The random variable X assumes a geometric distribution. And the expected value $E(X)_{\mathbf{P}}$ is calculated from the default population distributions $\mathbf{P_s}$, $\mathbf{P_a}$,$\mathbf{P_d}$ on the sex, age group, and domicile region attributes inside the biometric identity database.

We first introduce the required variables in modeling the identity tracing procedure during the biometric-identity matching procedure. The deduction of (1) takes place in section 2.3.

2.1 Database Related and Search Variables

The database related variables and the search variable are discrete random variables. They are summarized in Table 1.

The random variable S represents the sex of an identity. It takes on two values 1 (male), and 2 (female) with the corresponding probabilities $\{p_1^s, p_2^s\}$. The random variable A represents the age group to which an identity belongs. In this paper, A has 8 possible values with the corresponding probabilities $\{p_1^a, p_2^a, ..., p_8^a\}$. The values represent 8 different age groups that covers the age span of an entire population. The values $\{p_1^a, p_2^a, ..., p_8^a\}$ are taken from the current census data as shown in table 2.[1]

[1] National Institute of Statistics and Economics Studies, `http://www.insee.fr/en/themes/tableau.asp`

Table 1. Random Variables

Random Variable	Description
S	Sex: 1 (M), 2 (F)
A	Age Group: $1, 2, ..., 8$
D	Domicile Region: $1, 2, ..., 101$
X	The number of searches needed before the condition $S = i$ **and** $A = j$ **and** $D = k$ is satisfied, for fixed i, j, k

Table 2. Age Distribution

Age Group	Probability
< 15 ans	0.185
$15 - 24$ ans	0.123
$25 - 34$ ans	0.123
$35 - 44$ ans	0.134
$45 - 54$ ans	0.136
$55 - 64$ ans	0.127
$65 - 74$ ans	0.081
> 75 ans	0.09

The random variable D represents the domicile of an identity. This variable has a range of 1 to 101 with the corresponding probabilities $\{p_1^d, p_2^d, ..., p_{101}^d\}$. This is a French scenario because currently there exists 101 French departments.[2] The random variable X represents the number of sequential searches required before one hits a personal record in the identity database that satisfies the specified sex, age group, and domicile region criteria. The search random variable X has a value range from 1 to the size of a drawer TT, and it assumes a discrete geometric distribution.

2.2 Anonymity Related Variables

The variables described in this section are related to the quantization of anonymity levels for a specific criterion. To guarantee the k-anonymity requirement for a specific criterion, we require that each specified criterion in the released table has at least k occurrences [3]. Table 3 summarizes the variables.

$Tol_{s,a,d}$**:** the minimum standard needed for achieving indistinguishability among people who satisfy the respective criterion. In order to achieve k-anonymity, one should have $Tol_s \geq 2k, Tol_a \geq 8k, Tol_d \geq 101k$.

ω_i**:** the weights associated with the variables Tol_s, Tol_a and Tol_d respectively. They represent the reliability of information held by an investigator. The weight ω_i is normally set at 1 unless there are reliability issues on the information, which would lead ω_i to a value strictly less than 1. The value $\omega_i = 0$ indicates that the specific criterion is not used in the investigation.

Tol**:** the global value that represents the tolerance for all the criteria used in the investigation. It is defined as

$$Tol = \omega_1 Tol_s + \omega_2 Tol_a + \omega_3 Tol_d. \tag{2}$$

We explain below the lower bound associated with each anonymity related Tols (see the variable description above). In the course of sequential searching for a

[2] `http://www.insee.fr/en/methodes/default.asp?page=definitions/departement.htm`

Table 3. Anonymity Related Variables

Variable	Description
k	the anonymity-preserving level k
Tol_s	the minimum standard needed for achieving indistinguishability among people who have the same sex criterion
Tol_a	the minimum standard needed for achieving indistinguishability among people who have the same age group criterion
Tol_d	the minimum standard needed for achieving indistinguishability among people who have the same domicile criterion
ω_1	weight for Tol_s, $0 \leq \omega_1 \leq 1$
ω_2	weight for Tol_a, $0 \leq \omega_2 \leq 1$
ω_3	weight for Tol_d, $0 \leq \omega_3 \leq 1$
Tol	global tolerance $:= \omega_1 Tol_s + \omega_2 Tol_a + \omega_3 Tol_d$

specific criterion, we use a random variable Y to represent the outcome of each search, whether be success or failure. To preserve k-anonymity, we require that the expected number of successful identifications exceeds k among n sequential searches. That is

$$E(Y) = np \geq k \Longleftrightarrow n \geq \frac{k}{p} \tag{3}$$

where n is the number of searches, p the probability of a successful identification. The value p is determined by the probability associated with the particular search criterion. For practical calculations as used in deriving the lower bounds associated with Tol_s, Tol_a, Tol_d, we have used a rough estimate $p = 1/\#Group$, where $\#Group$ is the number of groups under a specific criterion.

2.3 Derivation of Formula (1)

We prove the privacy-preserving drawer size formula (1) in this section. To determine the drawer size TT, one must fix a priori a search scenario specifying the sex, age group, and domicile region criteria. For instance, the search profile contains the following information: sex $S = i$, age group $A = j$, domicile $D = k$, where i, j, k are fixed in their respective domains. Henceforth, the expected value $E(X)_{\mathbf{P}}$ based on the default distributions $\mathbf{P_s}$, $\mathbf{P_a}$,$\mathbf{P_d}$ is

$$\begin{aligned} E(X)_{\mathbf{P}} &= \frac{1}{Pr(S = i; A = j; D = k)} \\ &= \frac{1}{p_i^s \cdot p_j^a \cdot p_k^d} \end{aligned}$$

assuming a geometric distribution on the search random variable X. Thus, the number of persons in the drawer satisfying the condition $(S = i; A = j; D = k)$ is expected to be about $TT/E(X)$. To preserve indistinguishability among the matched individuals, $TT/E(X)$ must exceed the anonymity tolerance $Tol(k)$:

$$\frac{TT}{E(X)} \geq Tol(k) \Longleftrightarrow TT \geq E(X) \cdot Tol(k).$$

The formula (1) now follows if we let $TT = E(X) \cdot Tol(k)$. And one may obviously without loss of generality assume $N_{expire} = 0$ in the proof.

3 Empirical Study of System Reliability

3.1 Attack Model

We present an attack model here, based on which the reliability of the system can be defined. In an identity theft scenario, the attacker assumes the identity of another person (or he hijacks the biometric data of another person). Presenting himself before the registration authority, the theoretical probability of his success in finding a match between the drawer ID of the identity and the drawer ID of the biometric data is:

$$p = \frac{1}{NT} \tag{4}$$

and we define the system reliability as $1 - p$. The lower attack probability is equivalent to a higher system reliability.

3.2 Test Methodology

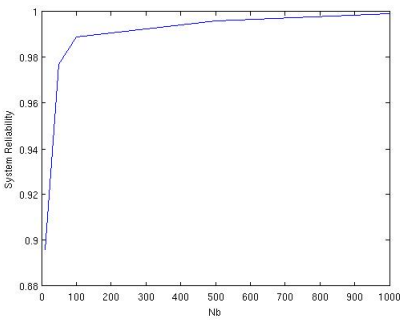
Our confidence tests are based upon the attack model described above. The test is performed on the prototype setbase system. Specifically for a fixed population of 10,000 and a fixed number of drawers, we perform identity thefts at the registration authority, and record the number of attack successes. The test results are recorded in Table 4. The theoretical attack rates are recorded in the second column. The observed attack rates are recorded in the third column. The observed attack rate is expressed as a ratio of the number of observed attack successes over the number of attack attempts. For the statistics generated in this section, we have performed attack attempts in the range of 1000 to 5000. The system reliability rates are obtained by subtracting the observed attack rates from 1. Figure 5 is a plot of the system reliability versus the number of drawers NT.

The observed attack rates adhere to the theoretical attack rates. This indicates that the basic architecture of the prototype system is sound and corresponds to how it should be. The source of absolute deviation between the theoretical rates and observed rates stems from the system performance fluctuations, and the inherent system errors.

Table 4. Probability of Attack Success for various NT

Test Population = 10,000		
Nb. Drawers NT	Theoretical attack Prob.	Observed Prob.
10	10%	10.4%
50	2%	2.3%
100	1%	1.1%
500	0.2%	0.4%
1,000	0.1%	0.1%
5,000	0.02%	0.01%

Table 5. Confidence Test: System Reliability versus NT



4 Privacy versus Security

For sufficiently large population, the variables number of drawers (NT) and size of a drawer (TT) are related by

$$Population \approx NT \cdot TT. \tag{5}$$

If we need privacy protection in the system, one possible solution would be to determine the minimum drawer size by means of formula (1). We then may adjust the number of drawers to reach a suitable system reliability level. We make this analysis rigorous in this section. Without consideration of privacy, we could chose a suitable NT based on the test results described in the previous section (Table 4).

We need the following data for carrying out the relational analysis: TT, NT, anonymity level k, attack probability, system reliability. Precisely, we first calculate the anonymity-preserving drawer size TT based on the general formula (1). The next step is to calculate NT using the formula (5), and the corresponding attack probability is calculated using the single attempt attack model (4). The reliability of the system is subsequently derived by subtracting the attack probability from 1.

We demonstrate the relation between TT and NT by studying the relation between the anonymity level variable k and the system reliability variable. Figure 1 and Figure 2 are two typical relational plots for the variables k and system reliability. Figure 1 reflects the entire range of $k, 2 \leq k \leq 30$. Figure 2 reflects the range $k, 2 \leq k \leq 10$ within which the highest system reliability rates occur. As can be been from the plots, the relation between k and system reliability is linear, inversely proportional. This fact again illustrates the underlying principle of the setbase: a stricter anonymity requirement (large k) requires a larger drawer size to be in place, that corresponds to a decrease on the number of drawers which in turn leads to a lower system reliability rate.

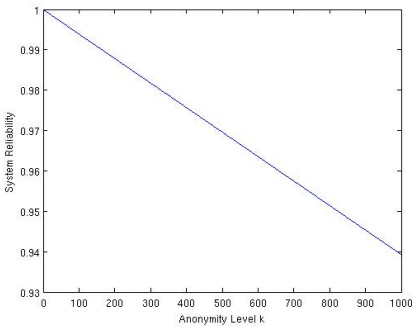


Fig. 1. System Reliability vs. Anonymity Level k, $2 \leq k \leq 1000$

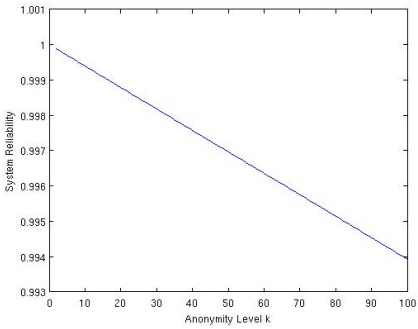


Fig. 2. System Reliability vs. Anonymity Level k, $2 \leq k \leq 100$

5 Conclusion

We have in this paper proposed a privacy-preserving drawer size standard for the setbase. The standard incorporates well known data privacy metrics such as k-anonymity and l-diversity as part of the drawer size formulation. The mathematical formulation gives one the ability to adjust the drawer size according the desired level of privacy. The future research in this direction includes issues such as how one can incorporate other database privacy notions into the privacy-preserving drawer size standard.

Acknowledgments. We would like to thank Vincent Pineau for implementing the prototype setbase in JAVA.

References

1. Assemblée Nationale No. 3599. Rapport relative à la protection de l'identité. Technical report (2011)
2. Didier, B., Rieul, F.: Person identification control method and system for implementing same, Patent US7724924 (2010)
3. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. Knowl. Data Eng. 13(6), 1010–1027 (2001)
4. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information (abstract). In: PODS, p. 188 (1998)
5. No.126 SÉNAT. Proposition De Loi: relative à la protection de l'identité. Technical report (2011)
6. Shamir, A.: Adding privacy to biometric databases: The setbase approach. Presentation Slide (2009)