Collecting Data while Preserving Individuals' Privacy: A Case Study

Alexis Bonnecaze [1], Robert Rolland [2]

Aix Marseille University
IML and ERISCS
13288 MARSEILLE Cedex 9, France
[1] alexis.bonnecaze@univ-amu.fr
[2] robert.rolland@acrypta.fr

***Abstract*—Several companies exploit medical data to better understand medication consumption patterns. Their analyses are useful to various health actors in order to enhance health care management. In this article, we focus on a configuration which allows a network of pharmacies to forward medical data to a private company in order to construct a database. Pharmacies must operate in full compliance with legal requirements in terms of confidentiality and privacy. We show that our solution fulfills all the requirements. Our work leads us to introduce the concept of generalized discrete logarithm problem which is proven to be as hard as the discrete logarithm problem.**

***Index Terms*—Privacy, Hash function, ElGamal ciphering**

I. Introduction

With the development of the digital world, a growing number of data are created every day in our society. These data can be very useful in many fields such as for example, commerce, marketing or medicine. They have a market value and are likely to be sold or to be made available to organizations or companies specialized in data analysis. However, these data often contain sensitive information that should not be leaked. Thus, data should be pre-treated in order to eliminate records which are to remain secret while preserving the consistency of the data. Moreover, statistical analysis of the data should not lead to the knowledge of any individual information. As an example, let us consider a database containing customer names with their purchases. It is possible to deduce clients' profiles from statistical analysis. In order to insure privacy, clients' name should be erased from the records but at the same time, it is required to be able to detect that two distinct articles have been purchased by the same client.

In this article, we focus on a real case study which addresses this problematic in the particular case of medical field. Indeed, our solution could be similarly applied to many other fields. We consider here a company which collects data from a number of pharmacies for statistical or economic analyses. In particular, these data contain patients' names and information regarding the patients like the name of drugs they have bought.

The paper is organized as follows. Next section gives an overview of the chosen case study, including the different actors and the requirements attached to them. Section III addresses the problem of anonymizing a pharmacy which sends a data and Section IV presents our protocol to anonymize the identity of the patients and fulfill the requirements given in Section II. Section V is devoted to security considerations.

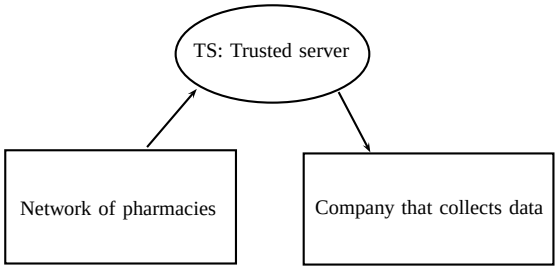
II. The case study

A company collects medical records from a set of pharmacies in order to create a database. Each medical record has two parts: a header containing information related to the identity of the patient and a body which contains various medical data. The field format of the header is the same for all the pharmacies. In order to comply with the law, pharmacies shall be under an obligation not to disclose information of the header. The body part of the record can be made public if it cannot be related to the identity defined in the header, whereas the header should be blinded. Remark that collecting data without headers does not constitute a solution since it fails to recognize whether two records involve the same patient or not. We therefore need to blind the headers while enabling to detect when two headers are identical. In an architecture point of view, it is technically possible to avoid the use

of a trusted party. However, the administration (in our case the CNIL department in France) requires to use a trusted third party, called TS, to avoid a direct contact between pharmacies and the company, and to insure that the protocol is fairly applied and that the cryptographic material is well-managed. Thus, we must consider that the existence of TS is a constraint of the problem. Therefore, having taken into account all the aforementioned requirements, pharmacies acquire the administrative rights to get the data out.

As illustrated in Figure 1, the architecture of the system is a network with three components: the network of pharmacies, the trusted server (TS) and the company which collects medical records.

Fig. 1: The network contains three components.



The objective of the system is to enable a pharmacy to forward a data to the company with the following requirements:

1) individual privacy must be preserved,
2) two records involving the same patient must have the same header.

Each component has the following requirements. Pharmacies must use a tamper proof box which makes transparent to the pharmacist the process of transmission of data. This box is able to encrypt using ElGamal algorithm. Key management is achieved by a certification authority. It is assumed that each box knows the public key of all the other boxes and that of the company. The trusted server needs to know the signature public key of the group of boxes. It is trusted in regards to transmission and non-disclosure of data transmission. However, it is not entitle to manage sensitive information. Its main work is to forward data to the company after having blinding it using a random number. In order to enhance privacy, TS is not authorized to know which pharmacy sends the data. Thus, TS should not be able to link any data with a pharmacy.

In terms of network transmission, it is assumed that a pharmacy can reach, through its box, TS and any other pharmacy, and TS can reach the company. The network of boxes is managed in a centralized way.

A. Cryptographic concerns

Our protocol makes use of cryptographic primitives. Every encryption uses elliptic curve ElGamal encryption. We introduce here the mathematical objects that we will use thereafter. Let us consider a cryptographic elliptic curve Γ over a prime field $\mathbb{F}_p$. Let Γ_p be the set of $\mathbb{F}_p$-rational points of Γ and $n = \#\Gamma_p$ the number of $\mathbb{F}_p$-rational points of Γ. We suppose that Γ is such that n is a prime number. Let us denote by G the cyclic group of order n of rational points on Γ and let P be a public generator of G. The curve Γ is chosen in order that the discrete logarithm problem be hard. Examples of such curves can be found in [5], [6] or [2]. Let H be a public map-to-point function as defined for example in [7] or [4]. Namely H transforms a message m to a point $H(M)$ of the curve and acts as a hash function.

III. Anonymization of the pharmacies

The set of pharmacies represents a private network. For this network, we choose to use a well known onion routing technique [8]. Each box of a pharmacy represents a node. In order to transmit a message to TS, a box has to

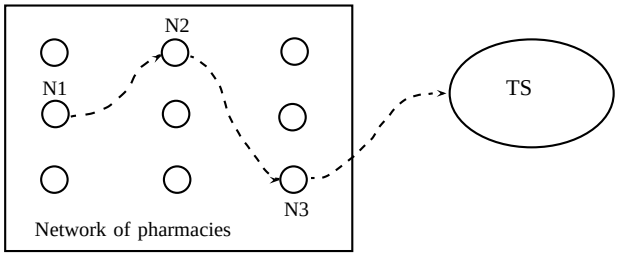
1) randomly choose a set of ordered nodes to provide a circuit $(N_1,\ N_2, \dots N_t)$ through which the message m will be transmitted.
2) Encrypt the message m with the public key K_t of N_t, then encrypt the result with the public key of N_{t-1} and so on, until obtaining

$$C = E_{K_1}(E_{K_2}(\cdots(E_{K_t}(m))\cdots)),$$

where E_{K_i} is the encryption function using the public key K_i. At each level, the box includes information regarding identity of the next node to which the onion must be transmitted.
3) As the onion C passes to each node in the circuit, a layer of encryption is peeled away by the receiving node. Decryption is performed using the private key corresponding to the public key with which the layer was encrypted.
4) The last node N_t transmits the original message m to TS.

Onion routing technique is used to hide the identity of the box which sends the message. Indeed, TS only knows the identity of N_t. We will see later that part of the original message is also encrypted in order to fulfil the aforementioned requirements.

Fig. 2: A random circuit is constructed to anonymize the sender when sending its message to TS.

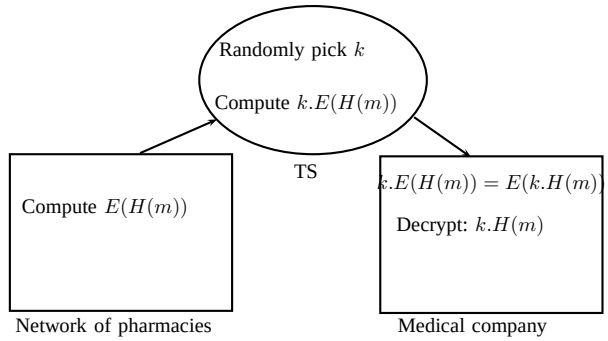


IV. Anonymization of the header

In this section, we describe the cryptographic protocol which allows the header m to be anonymized. First, let us consider the following simple solution: the header m of a record is hashed using a hash function. This solution is far from being secure, as it is vulnerable to dictionary attack. It is therefore necessary to provide a more comprehensive mechanism.

An overview of our protocol is illustrated in Figure 3, where the encryption function is denoted $E()$. The important property of this function is that $k.E(M) = E(k.M)$ for any message M. This property ensures the second requirement, the encryption ensures the first requirement with regard to TS and the masking by k ensures the first requirement with regard to the company.

Fig. 3: Anonymization of the record.



We describe here in more details the process to anonymize the header m of the record to transmit.

The cryptographic set up phase is as follows:

1) The trusted party TS picks at random a key k such that $0 \leq k \leq n-1$ and keeps it secret.
2) The company picks at random a key a such that $0 \leq a \leq n-1$ and keeps it secret. Moreover
3) the company computes the point $Q = aP$ and transmits it to the network of pharmacies (this is the public key of the company).

When the set up is done, any pharmacy's box can forward a data.

1) A box P draws at random an integer k_1 between 0 and $n-1$. Then P computes

$$P_1 = k_1 P \quad P_2 = H(m) + k_1 Q.$$

The points P_1 and P_2 are sent to the trusted third party TS.

2) The trusted third party TS computes, using its secret key k, the two following points

$$R_1 = kP_1 \quad R_2 = kP_2$$

and sends R_1 and R_2 to the company.

3) Now the company computes the anonymous number AN associated to the header

$$AN = (R_2 - aR_1)_x$$

where $(R_2 - aR_1)_x$ denotes the x-coordinate of the point $R_2 - aR_1$.

Remark 1: The random number k_1 drawn by the pharmacy's box must be recalculated for each record. However the secret key k of the trusted third party must remain the same throughout the study.

Proposition 1: The anonymous number AN is

$$AN = \big(kH(m)\big)_x.$$

Proof: We compute $R_2 - aR_1$ and obtain successively:

$$R_2 - aR_1 = kH(m) + kk_1Q - akP_1$$

$$= kH(m) + kk_1aP - akk_1P = kH(m).$$

■

V. Security considerations

In this section, we show that our protocol fulfills the security requirements. This system does not intend to prevent any pharmacy to be corrupted. Indeed, it is technically impossible to prevent a pharmacist to disclose information he or she has access.

We consider security in regards to the different actors. The security from a box to TS lies to DDH problem and the security from TS to the company lies to the generalized discrete logarithm problem introduced in Section V-B and analyzed in Section V-C.

A. Privacy in regards to TS

It is required that TS be able to authenticate a message received from a pharmacy without being able to distinguish what pharmacy sent it. Every message is dated and signed using the boxes' private key (all the boxes use the same key). At the transport level, the onion routing method provides anonymization with regard to TS. At this level, confidentiality is not mandatory since the header of the data is encrypted by an elliptic curve Elgamal ciphering. More precisely, the header $H(m)$ is masked by k_1Q where k_1 is random.

Proposition 2: Under the assumption that DDH problem is hard on the group of the chosen curve, the trusted party is not able to distinguish whether two encrypted headers represent the same plaintext header or not.

Proof: It is well known that this type of ciphering is indistinguishable under chosen plaintext attack (IND-CPA) in the random oracle model, as far as we work on a group where the decisional Diffie-Hellman problem is hard (see [9] or [1]). In particular, it means that TS is not able to distinguish whether two encrypted headers represent the same plaintext header or not. ■

B. Privacy in regards to the company

The company needs to be sure that the sender is TS. Authentication is done by adding a timestamp and signing the message. Even-though confidentiality is not required, a protocol like TLS may be used. When the company receives the plaintext message, it remains to treat the header. Using homomorphic properties of the ciphering, the company can eliminate the mask k_1. The header is now protected by the blinding factor k. The underlying security problem is the generalized discrete logarithm problem on the chosen elliptic curve. This problem which, as far as we know, has not been considered yet in the literature is analyzed in the next subsection.

C. Generalized discrete logarithm of order s

It remains to study the following problem. Suppose that an attacker knows some identities of clients of the network of pharmacies and the set of corresponding blinded headers. Since the blinding value k is fixed, is he able to calculate k?
Mathematically, this problem can be written as follows. Let an integer s be such that $1 \leq s \leq n-1$, let a (non ordered) set of rational points $A = \{A_1, \ldots, A_s\}$ and let k be an integer such that $1 < k < n-1$. We denote kA the set $\{kA_1, kA_2, \cdots, kA_s\}$. The problem P_s of the generalized discrete logarithm of order s on the group Γ_p is the following: Given A and $A' = kA$, calculate k.

Remark 2: The knowledge of A and $A' = kA$ is equivalent to the knowledge of $B = \complement A$ and $B' = kB = \complement A'$. In particular, P_{n-1} is equivalent to P_1, the discrete logarithm problem (DLP).
In our case study, the value s is much smaller than n and in practice, we may assume that $500 \leq s \leq 10^6$. We will show that P_s is at least as hard as DLP.

Theorem 1: Suppose we know an algorithm $\mathcal{A}(\Gamma_p, s)$ which solves P_s in a time bounded by $T(s)$, then it is possible to construct an algorithm which solves DLP on Γ_p in a time bounded by $T(s) + st_0$ where t_0 is the time needed to choose an integer m and to calculate two scalar multiplications on Γ_p.

Proof: Let A_1, $A'_1 = kA_1$ be an instance of the DLP. Let us choose distinct integers $m_2, \cdots, m_s$ such that $1 < m_i < n$ in order to construct the points $A_i = m_iA_1$ and $A'_i = m_iA'_1$. We have $A'_i = m_ikA_1 = km_iA_1 = kA_i$. Thus, if $A' := \{A'_1, A'_2, \cdots, A'_s\}$, we obtain $A' = kA$. By this way, we just constructed an instance of P_s. The time needed for this construction is bounded by st_0. Applying the algorithm $\mathcal{A}(\Gamma_p, s)$ to this instance of P_s, we can obtain k. We have therefore solved DLP in a time bounded by $T(s) + st_0$. ■
Consequently, if we had a practical algorithm to solve P_s, s being sufficiently small (in order that st_0 can be reached in practice), then we could solve DLP over Γ_p. As an example, if we choose a curve over $\mathbb{Z}/p\mathbb{Z}$ where the size of p is around 256 bits, then from Weil's bound, the size of n is of the same order. This means that n is of order 2^{256} and the best known algorithms to solve DLP need about 2^{128} operations. If s is bounded by 10^6 (our case study), then s is negligible compared with 2^{128}. Thus, unless breaking the DLP for this size, we cannot obtain an algorithm to solve P_s with a number of operations significantly less than 2^{128}.

VI. Conclusion

This article solves a problem which has effectively been encountered in an industrial framework. Since the company cannot directly reach the pharmacies and receives data via the box and TS, database privacy techniques like differential privacy [3] are not adequate. Our protocol has a wide range of applications since statistical analyses of data are used extensively and privacy is becoming a major concern. We showed that the solution fulfills all the requirements regarding privacy concerns. Moreover, the company is able to distinguish whether two records involve the same patient while this property

is not allowed to the third party designed to forward the data and blind the header. Our analysis led us to introduce the concept of generalized discrete logarithm problem of order s and we proved that this problem is at least as hard as the discrete logarithm problem.

REFERENCES

[1] Pierre Barthélemy, Robert Rolland and Pascal Véron. Cryptographie principes et mises en oeuvre- 2e édition. *Editions Hermes Lavoisier, Lavoisier.* 2012.

[2] ECC Brainpool. ECC Brainpool Standard Curves and Curve Generation, October 2005. http://www.ecc-brainpool.org/download/Domain-parameters.pdf.

[3] Cynthia Dwork Differential privacy: A survey of results. *In Theory and Applications of Models of Computation, Springer*, pages 1–19, 2008.

[4] Thomas Icart. How to Hash into Elliptic Curves. *CRYPTO 2009: 303-316*, 2009.

[5] Hamish Ivey-Law and Robert Rolland. Constructing a database of cryptographically strong elliptic curves. *Proceeding of SAR-SSI*, , 2010. http://www.acrypta.com/index.php/telechargements#ARCANA.

[6] NIST. NIST-FIPS 186-3 (website), 2009. http://csrc.nist.gov/publications/fips/fips186-3/fips_186-3.pdf.

[7] Dimitrios Poulakis and Robert Rolland. A signature scheme based on elliptic curve discrete logarithm and factoring. *Cryptology ePrint Archive, 2012/134*, 2012.

[8] Michael Reed, David Goldschlag and Paul Syverson. Hiding routing information. *Lecture Notes in Computer Science*, 1174:137–150, 1996.

[9] Yiannis Tsiounis and Moti Yung. On the security of Elgamal based encryption. In *Proceedings of the First International Workshop on Practice and Theory in Public Key Cryptography: Public Key Cryptography*, PKC '98, pages 117–134, London, UK, UK, 1998. Springer-Verlag.