Multi-centric Universal Pseudonymisation for Secondary Use of the EHR

Luigi Lo Iacono
C&C Research Laboratories, NEC Europe Ltd., Germany

Abstract. This paper discusses the importance of protecting the privacy of patient data kept in an Electronic Health Record (EHR) in the case, where it leaves the control- and protection-sphere of the health care realm for secondary uses such as clinical or epidemiological research projects, health care research, assessment of treatment quality or economic assessments. The paper focuses on multi-centric studies, where various data sources are linked together using Grid technologies. It introduces a pseudonymisation system which enables a multi-centric universal pseudonymisation, meaning that a patient's identity will result in the same pseudonym, regardless of which participating study center the patient data is collected.

Keywords. Pseudonymisation, Privacy, Multi-centric Studies, HealthGrid

1. Introduction

The Electronic Health Record (EHR) is a personal medical record in digital format, containing in the first instance information relating to the current and historical health, medical conditions and medical tests of its subject. It is primarily used in the treatment context in which the patient's identity data is needed and protected by medical secrecy. But the EHR also serves as a basis for other purposes denoted as secondary uses, such as clinical or epidemiological research projects, health care research, assessment of treatment quality or health economy.

Characteristic for the secondary use is, that the patient data leaves the control- and protection-sphere of the medical secrecy. Patient data is very sensitive personal data. To make it available outside of the treatment context, various legal and ethical aspects have to be considered [1]. First of all, the patient has to agree to participate with his patient data to a secondary use by giving an informed consent. Then the Personal Identifiable Information (PII) of the patient has to be removed from the EHR before the use in secondary contexts is allowed. Thus, in such a context the use of the EHR is allowed after anonymisation and must be performed whenever possible, but although the identity of the patient does not matter in secondary contexts, it is, however, not always possible to simply anonymise the EHR. In many scenarios of secondary use, the correct association between a single patient and his EHR from distinct sources or distinct points in time is essential. Examples are the provisioning of follow-up data at a later point in time, the withdraw of samples or data on specific patient's request or the quality control of the data such as the checking for double-booking. This usually prevents the application of anonymisation and demands instead for pseudonymisation schemes. In some scenarios even the re-identification of a patient is required, when clinically relevant information arises during the course of a study which might have a direct impact on the treatment of a patient. In such cases, ethical principles demand recontacting and informing all relevant patients about the findings.

In summary, depending on the kind of research network and its requirements, distinct procedures for anonymisation or pseudonymisation are appropriate. In this paper we focus on research networks where various data sources must be linked together, such as given by a multi-centric study that uses data from EHRs, but also data or samples from biomaterial banks, or follow-up data at a later point in time and where the possible recontact of patients is required. Such requirements are common for studies as carried out for example by the @neurist project[1] funded within the European IST FP6 program. The @neurist project aims to integrate data which spans all length scales, from molecular, through cellular to tissue, organ and patient representations in order to develop advanced decision-support systems to help treat cerebral aneurysms. These data are increasingly heterogeneous in form, including textual, image and other symbolic structures, and are also diverse in context, from global guidelines based on the broadest epidemiological studies, through knowledge gained from disease-specific scientific studies, to patient-specific data from electronic health records.

2. Previous Work

Available pseudonymisation systems can roughly be classified depending on their ability to be reversible or not.

One-way pseudonymisation systems generate pseudonyms in such a way that it is almost impossible to re-identify the patients from the generated pseudonyms. Keyed cryptographic one-way functions [2] are a common technology to implement this property (also referred to as one-way encryption in this context). The pseudonyms could be produced either at clinical centers or through a central Trusted Third Party (TTP) service. For multi-centric studies this step has to be performed by an TTP in order to obtain pseudonyms which are universal to the study, creating links through all information on one patient. An essential prerequisite here is some form of PII, which identifies the patients unambiguously. New information on the same patient would produce the same pseudonym and therefore enable updates to the data. The main advantage to this system is the high level of privacy protection it provides. From the viewpoint of recontacting of patients, the natural disadvantage is that it would be very difficult to trace back to original patients. The conceptually simplest way to enable re-identification in such systems would be to store the association between PII and pseudonyms at the TTP. Such a mapping database is, however, an attractive target for attackers and violates the medical secrecy [3].

Reversible pseudonymisation systems allow the patient to be re-identified through cryptographic mechanisms and the existence of cryptographic keys in particular. The re-identification through decryption of the pseudonym eliminates the need of maintaining a mapping database. In the case of symmetric cryptosystems – which uses one and the same key for encryption and decryption –, one entity holds the key to the pseudonym, creating a level of security very much like the patient/clinician confidentiality relationship. If the clinical centre holds such a key, they are able to re-identify participating patients. A solution to ensure that re-identification of individuals

[1] http://www.aneurist.org

is strictly controlled is to use asymmetric cryptosystems instead of symmetric ones, where the public encryption key is used by the pseudonymisation service, but the private decryption key is only known to and only in possession of the study's medical and ethical advisory board. Such an approach is to the knowledge of the author, however, neither widely used nor discussed. Instead, more commonly an extra pseudonymisation stage is added to the scheme based on symmetric cryptography including a corresponding separate secret key. The second pseudonymisation step of the so-called dual-pass pseudonymisation systems is introduced in such a way that the two pseudonymisation procedures are independent from and do not know anything about each other. This “additional privacy safeguard” allows for much stricter control over linking research information back to a patient, and the situation when the two pseudonyms are brought together can be regulated by strict operating procedures. In single-centric studies the first pseudonymisation step can be performed within the study center [4]. For multi-centric studies, however, both have to be run by distinct TTP in order to obtain the correct linkage between data from different sources. LIPA [5] is an example for such an architecture and the only one the author is aware of. It relies on asymmetric cryptographic techniques and two distinct TTP in order to calculate an universal pseudonym. However, the first TTP receives the patient's National Health Service (NHS) number which reveals PII to the TTP since the NHS number is the common and unique identifier for patients in England and Wales.

For the depicted and targeted research network type and according to the available technologies a reversible dual-pass pseudonymisation system is most desirable. However, to generate universal pseudonym which can link all the diverse data sources on one patient together, the known approaches rely either on a central TTP service which has to obtain at least some form of PII of the patient or are simply not capable to generate such universal pseudonym.

3. Multi-centric Universal Pseudonymisation

To overcome the limitations, that a patient has to be treated always in the same study center or that parts of his PII has to be sent to a central pseudonymisation service in order to obtain the corresponding pseudonym within a particular research network, the following pseudonymisation scheme is proposed. It enables a multi-centric universal pseudonymisation, meaning that a patient's identity will result in the same pseudonym, regardless in which participating study center the patient data is collected. Furthermore, the study centers do not have to reveal the patient's identity data to an TTP service in order to retrieve the corresponding inter-clinic and unambiguous pseudonym. Instead, the first pseudonymisation step is performed locally and a trusted pseudonymisation service then performs a second pseudonymisation step in which the inter-clinic and unique pseudonym is computed based on the output of the locally performed pseudonymisation which does not contain any PII.

The proposed scheme relies on a number theoretic problem usually used for constructing asymmetric cryptosystems. More specifically the Discrete Logarithm Problem (DLP) is used which is the basis for asymmetric cryptographic schemes such as the Diffie-Hellman (DH) [6] key agreement and the ElGamal [7] public key encryption and signature scheme. As for DH and ElGamal, the proposed scheme can be implemented in any group where the DLP is infeasible, including e.g. the group of an elliptic curve defined over a finite field which forms the fundament of Elliptic Curve

Cryptography (ECC) [8, 9]. For the sake of simplicity and clarity, however, the following descriptions are focused on the multiplicative group of a finite field $\boldsymbol{Z}_p$ (where p is prime) only.

The cryptographic techniques used do not allow the reversal of the pseudonymisation to a patient's identity by decrypting the pseudonym. The reverse mapping has to be performed by a corresponding database maintaining the associations between the patient's identifier and the universal pseudonym.

3.1. Symbols and Abbreviations

The following symbols and abbreviations are used to increase the clarity of the paper.

E	Set of all patients.
S	Set of all study centers.
sk^s	Secret key of key pair of study center $s \in S$.
sk^{TTP}	Secret key of key pair of trusted pseudonymisation service.
pk^s	Public key of key pair of study center $s \in S$.
pk^{TTP}	Public key of key pair of trusted pseudonymisation service.
PK	$PK = \{pk^s \mid \forall s \in S\} \cup \{pk^{TTP}\}$. Set of public keys of all study centers and the pseudonymisation service.
PK^s	$PK^s = PK \setminus \{pk^s\}$. Set of public keys of all study centers and the pseudonymisation service, except the public key of study center $s \in S$.
PII^e	Personal Identifiable Information of patient $e \in E$ which are used for pseudonym generation.
ID^e	An identification which identifies the patient $e \in E$ unambiguously and which is produced on its PII^e.
lid^e	Local pseudonym computed for the patient $e \in E$ based on its PII^e or ID^e respectively.
$lid^{e,s}$	lid^e computed by the study center $s \in S$.
gid^e	Global Pseudonym computed for the patient $e \in E$.
$rand[x, y]$	A random number function which chooses randomly a number from the interval defined by $[x, y]$.
p	A large prime number.
g	A primitive element modulo p of the multiplicative group of a finite field $\boldsymbol{Z}_p$ (also called generator).
OWF	An one-way function, such as cryptographic hash-functions like SHA-256 or Whirlpool.

3.2. Initialization Steps

In the initialization phase is based on the system-wide and public parameters p and g, which define a multiplicative group of a finite field $\boldsymbol{Z}_p$ with primitive element g corresponding to DLP-based cryptosystems. The parameters have to be chosen according to current key length proposal by the cryptographic community and the needs of the study [10]. Each participating study center then has to perform the following initialization steps:

1. Select a random number $r = rand[1, p\text{-}1]$.
2. Compute $\rho = g^r \; mod\, p$.
3. Distribute ρ to the trusted third party (the pseudonymisation service in charge of computing the inter-clinic pseudonym) and keep r secret.

The computed number ρ is the public key pk^s of study center s, the selected random number r is the private key sk^s. The TTP pseudonymisation service has to perform the same steps, but keeps both keys secret (see section 5).

Assuming the following setting for a subsequent example:

• Study Centre Alice:	$pk^A = \alpha,\ sk^A = a$
• Study Centre Bob:	$pk^B = \beta,\ sk^B = b$
• Study Centre Carol:	$pk^C = \chi,\ sk^C = c$
• Study Centre Dave:	$pk^D = \delta,\ sk^D = d$
• TTP Service Trent:	$pk^{TTP} = \tau,\ sk^{TTP} = t,\ \mathrm{PK} = \{\alpha, \beta, \chi, \delta, \tau\}$

After all system participants have performed the described initialization, the process to obtain a research network global and universal pseudonym is divided into two distinct phases. First, a local pseudonym is computed by the study center and second, the global pseudonym is computed by a pseudonymisation service as will be described in the following.

3.3. Computing a Local Pseudonym

The computation of a local pseudonym is composed of two sequential steps (see Figure 1). In the first step, the patient's PII used for pseudonym generation is transformed by the application of an one-way-function (*OWF*) in such a way,

- that the patient is identified unambiguously by the output of the *OWF*,
- that the variable input data is mapped to a fixed- and short-size output, and
- that the reverse operation is not feasible.

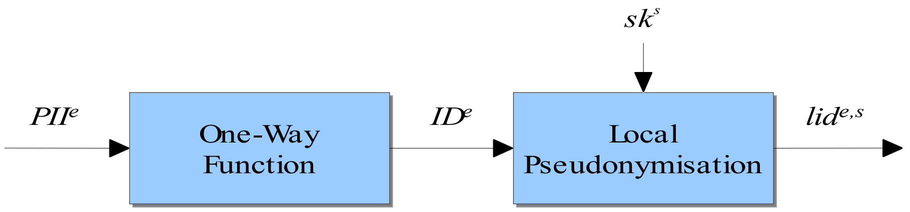

Figure 1. Study center computing a local pseudonym

Cryptographic hash-functions are a good candidate for such OWF. The output of the OWF (which in the case of hash-functions is the hash value) is denoted as $ID^e = OWF(PII^e)$. Note, that this step can be performed by any study center in the research network and will result in an identical ID^e for identical PII^e for a given patient $e \in E$. Thus, an essential prerequisite is the unique patient's PII.

From the generated ID^e the study center s can finally produce the local pseudonym in the second and last step by computing $lid^{e,s} = g^{sk^S + ID^e}$ *mod p*, using the research network related private key sk^S generated as described in section 3.2.

3.4. Computing the Universal Pseudonym

The global pseudonym gid^e for the patient $e \in E$ can be computed from the corresponding local pseudonymisation lid^e of the patient e by the central pseudonymisation service as follows: $gid^e = lid^{e,s} \cdot \prod PK^S \bmod p$.

Figure 2 illustrates the computation of the universal pseudonym based on the example setting introduced in section 3.2. If, for example, the study center Bob computes the local pseudonym $lid^{e,B}$ of patient e and sends it to the pseudonymisation service Trent, Trent is able to generate the universal pseudonym gid^e of patient e by performing the following computation: $gid^e = \alpha \cdot lid^{e,B} \cdot \chi \cdot \delta \cdot \tau$ *mod p*. Note, that regardless of which system participant (excluding Trent) computes $lid^{e,s}$, the global pseudonym gid^e will always be identical.

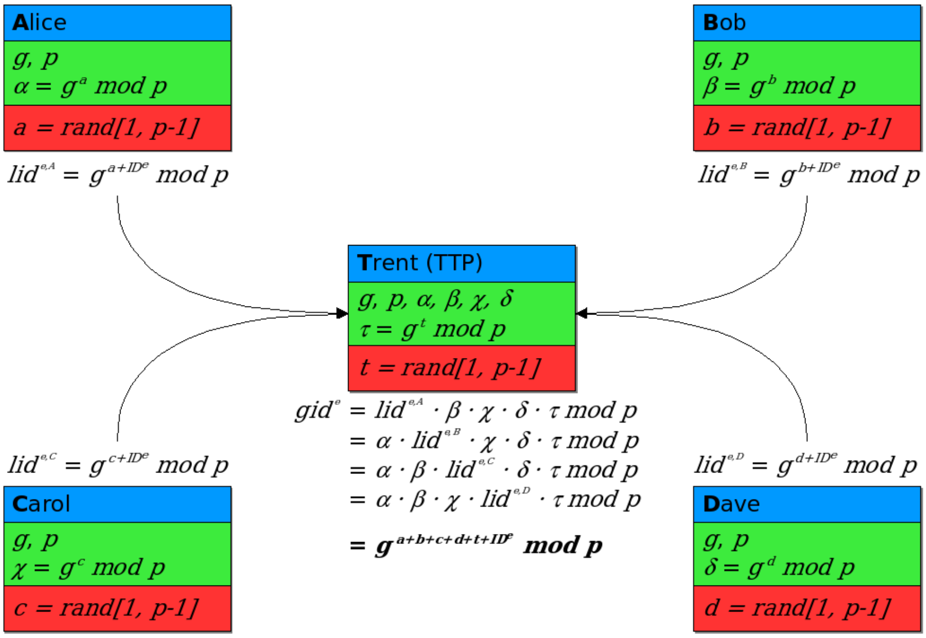
Alice
g, p
$\alpha = g^a \bmod p$
$a = rand[1, p-1]$
$lid^{e,A} = g^{a+ID^e} \bmod p$
Bob
g, p
$\beta = g^b \bmod p$
$b = rand[1, p-1]$
$lid^{e,B} = g^{b+ID^e} \bmod p$
Trent (TTP)
$g, p, \alpha, \beta, \chi, \delta$
$\tau = g^t \bmod p$
$t = rand[1, p-1]$
$gid^e = lid^{e,A} \cdot \beta \cdot \chi \cdot \delta \cdot \tau \bmod p$
$= \alpha \cdot lid^{e,B} \cdot \chi \cdot \delta \cdot \tau \bmod p$
$= \alpha \cdot \beta \cdot lid^{e,C} \cdot \delta \cdot \tau \bmod p$
$= \alpha \cdot \beta \cdot \chi \cdot lid^{e,D} \cdot \tau \bmod p$
$= g^{a+b+c+d+t+ID^e} \bmod p$
$lid^{e,C} = g^{c+ID^e} \bmod p$
Carol
g, p
$\chi = g^c \bmod p$
$c = rand[1, p-1]$
$lid^{e,D} = g^{d+ID^e} \bmod p$
Dave
g, p
$\delta = g^d \bmod p$
$d = rand[1, p-1]$

Figure 2. TTP computing a multi-centric universal pseudonym

3.5. Re-Identification

To re-identify the patient from a global pseudonym, mappings between global pseudonyms and local pseudonyms as well as between local pseudonyms and patient identifiers have to be maintained, since the proposed scheme does neither allow the retrieval of the local pseudonym from a particular global pseudonym nor the retrieval of the patient identifier from a local pseudonym.

The maintenance of these mappings can be realized in distinct ways. A common approach would be to keep the mapping between the patient identifiers and the local pseudonyms in the corresponding study centers and to store the mapping between the local pseudonyms and the global pseudonyms in the trusted pseudonymisation service. In such a scenario, to re-identify a patient, the ethical and medical board would first refer to the pseudonymisation service in order to retrieve the local pseudonym(s) for a given global pseudonym and would then ask all the study centers whether one of the local pseudonym(s) belongs to one of their patients.

4. Implementation and Evaluation

An implementation for evaluation purposes has been developed in Java. The pseudonymisation service provides a Web services interface to the study centers. The client side provides a user interface where the patient's PII necessary for the pseudonym generation can be inputted. From this PII the local pseudonym is computed and then transferred to the pseudonymisation service by calling the provided service. The response from the invoked pseudonymisation Web service includes the computed global pseudonym.

When, for example, the global pseudonym of Mr. Peter Patient born on January, 1st 1961 has to be generated, his ID^e is computed using the input data: 'Peter, Patient, 01-01-1961, m'. Since the ID^e is computed by for example hashing the input data in some suitable form, the ID^e is equal in each study center to `84c6fee073ffbe0854 049292888b845fdba1d25cb1afb29f593e9823bff0c6aa`, whereas the computed local pseudonyms differ from each other (all values are expressed in hexadecimal form):

- $lid^{e,A}$: local pseudonym computed by study center Alice
  `1e69473a1d59c7969c0c3482e6e855003c850e186e259c35a976`
  `8aef13e3fb6ab3ce6cfc8989c12daf6a7ccccaef1e03367c9e03`
  `3c0527ef4c312150e15650dc98878da34a43787c89819b23c6fd`
  `002755c9c01bbb7872599774da81881f9b78702462019185b761`
  `dae139a58c5ccd968aaa1df98bd7c148fe7ab2a1f883da42`
- $lid^{e,B}$: local pseudonym computed by study center Bob
  `2c86baf18b58b03743c213f325d0a4fb65cab1fb6b4bf9ca615e`
  `4cccfe7c2819ffdc24d2e6ca1266150a1e56bf07ea7e20674656`
  `131cfc52f13d1d6422b7bcf25fa2d4af4d33164b89a3d312be1f`
  `0c947bbe302df0fde58ce795c66371a3aafa05799ab0b30ff398`
  `db1993bb2386d23fa69d62881f61d08e4c21126b39314abb`
- $lid^{e,C}$: local pseudonym computed by study center Carol
  `0086391e496fdb8974ce2b9d64bf760d1c7152d9a323e9db30c6`
  `e80909d5e8ce4f56dd3d020a56801addd07b393e25b23221116c`

```
877fd5f6be370c70bc87de85b55be48f8a02f8ba445bcf36d113
443deb7430a735fc9f262eb4c418c8e3329136a4b495549d55de
cefab55f22a0e43b6f9149d823241c0ebd7bd62635043ec7b2
```

- $lid^{e,D}$: local Pseudonym computed by study center Dave

```
2c2baa66c47a79a529ff2bb21071a0f9929a58c1ab7672663c2b
83f49874b47a2fd12d1ffe7c8465f690372c6e2a2ee21cacba2a
05c8263e1b6b4fc528eec7004d9594e5f61d63fa5bfa2276307d
07a9a3a786c031f4377b54676b03c3c13824d29b6065e4197cc6
d3253e5b5dd1801726905d105878119428fb04acc7054da0
```

The global pseudonym computed by the pseudonymisation service out of the supplied local pseudonym and the appropriate public keys from the study centers, however, is again globally unique:

- gid^e: global pseudonym computed by pseudonymisation service Trent

```
98e1af507e6f33bf29fe5a1d67d4a05f626b57a5b030d794d85e
d302a0cdb4a1a1296d94d935496219bcc307fcce9a6c670edf03
df15093cc9b243e64e9efaba4571f2660ec2adfff482d463f45c
84abe0f1a5463676462bb3b032be6c00a0d0a14b071f11d90f4a
c1231e1bb514e2ee879f873100c73beaf5e70be0fe030fae
```

To reduce the size of the gid^e – which is bound to the cryptographic domain parameters –, it can e.g. be hashed using a cryptographic hash-function.

5. Security Considerations

Through the adoption of DLP-based asymmetric cryptography it is computationally infeasible to derive the random value r from the public domain parameters p and g as well as the public key r. The same holds for the computation of the local pseudonyms lid^e. Thus, nobody is able to obtain the ID^e from lid^e except the corresponding local pseudonymisation service.

In order to prevent anybody from computing the global pseudonyms gid^e from the local pseudonyms, the public key of the trusted pseudonymisation service must be kept secret.

The PII used to compute the local pseudonym must be choosen accordingly to render exhaustive search attacks infeasible.

Since the proposed pseudonymisation systems and its included cryptographic techniques are not reversible per se, corresponding mapping databases are required to re-identify a certain patient from its global pseudonym for matters discussed in section 1. Such components have to be safeguarded to prevent unauthorized access and henceforth the unauthorized re-identification of patients.

6. Conclusion and Outlook

The presented pseudonymisation system aims to contribute a possible solution to the problem of providing multi-centric health studies with universal pseudonyms enabling to cross-link the distributed but federated data sources of a study. It remains to be evaluated, however, whether the proposed system can support the real use cases of multi-centric studies in an appropriate manner. In this context the implications contained in the proposed system concerning for example membership changed during the course of a study as well as related issues such as re-keying and re-pseudonymisation. These aspects will be further investigated and analyzed within the @neurIST project.

Acknowledgment

This work was generated in the framework of the @neurIST Integrated Project, which is co-financed by the European Commission through the contract no. IST-027703.

The author would like to thank Frederik Armknecht, Jun Furukawa, Greg Kohring, Jun Wang, Kazue Sako and Obana Satoshi for many fruitful discussions as well as the anonymous referees for their constructive comments.

References

[1] J. Fingberg, M. Hansen, M. Hansen, H. Krasemann, L. Lo Iacono, T. Probst, and J. Wright: Integrating Data Custodians in eHealth Grids - A Digest of Security and Privacy Aspects, In: Christian Hochberger, and Rüdiger Liskowsky (Eds.), Informatik 2006 - Informatik für Menschen, Lecture Notes in Informatics (LNI) vol. P-93, pp. 695-701, 2006.

[2] S. Bakhtiari, R. Safavi-Naini, and J. Pieprzyk: Keyed hash functions, In: E. Dawson, and Jovan Golic (Eds.), Cryptography: Policy and Algorithms, Lecture Notes in Computer Science, vol. 1029, pp. 201-214, Springer-Verlag, 1996.

[3] K. Pommerening, and M. Reng: Secondary use of the EHR via pseudonymisation, In: L. Bos, S. Laxminarayan, A. Marsh (Eds.), Medical Care Compunetics 1, pp. 441-446, IOS Press, Amsterdam 2004.

[4] D. Kalra, P. Singleton, D. Ingram, J. Milan, J. MacKay, D. Detmer, and A. Rector: Security and confidentiality approach for the Clinical E-Science Framework (CLEF), In: Second UK E-Science All Hands Meeting, Nottingham (UK), 2003.

[5] N. Zhang, A. Rector, I. Buchan, Q. Shi, D. Kalra, J. Rogers, C. Goble, S. Walker, D. Ingram, and P. Singleton: A Linkable Identity Privacy Algorithm for HealthGrid, In: T. Solomonides, R. McClatchey, V. Breton, Y. Legré, and S. Nørager (Eds.), Proceedings of Healthgrid 2005, Studies in Health Technology and Informatics, vol. 112, 2005.

[6] W. Diffie, and M. E. Hellman: New directions in cryptography, In: IEEE Transactions on Information Theory, vol. IT-22:6, pp. 644-654, November 1976.

[7] T. El-Gamal: A public-key cryptosystem and a signature scheme based on discrete logarithms, In: IEEE Transactions on Information Theory, vol. IT-31:4, pp. 469-472, July 1985.

[8] V. Miller: Uses of elliptic curves modulo lage primes, In: Advances in Cryptology – Crypto’85, Springer Verlag, pp. 417-426, 1986.

[9] N. Koblitz: Elliptic curve cryptosystems, In: Math. Comp., vol. 48, pp. 203-209, 1987.

[10] A. K. Lenstra, and E. R. Verheul: Selecting Cryptographic Key Sizes, In: Journal of Cryptology, vol. 14:4, pp. 255-293, 2001.