Anonymization, tokenization, encryption How to recover unrecoverable data

O. DZIĘGIELEWSKA
olga.dziegielewska@wat.edu.pl

Military University of Technology, Faculty of Cybernetics
Urbanowicza Str. 2, 00-908 Wasaw, Poland

The data privacy is currently vastly commented topic among all the organizations which process personal data due to the introduction of the European Union's General Data Protection Regulation. Existing methods of data protection are believed to be sufficient as they meet the risk-based approach requirements in every mature organization, yet the number of publicly known data breaches confirms that this assumption is false. The aftermath of such incidents in countless cases prove that the risk-based approach failed as the reputational and financial consequences by far exceed the original estimations. This paper stressed the importance of the data layer protection from the planning, through design, until maintenance stages in the database lifecycle, as numerous attack vectors originating from the insider threat and targeting the data layer still sneak through unnoticed during the risk analysis phase.

Keywords: database lifecycle, inference attacks, data privacy, data breaches, GDPR.

1. Introduction

In the today's world, our personal data are processed by multitude of IT systems every day. We are unable to control the flow of such data from its creation in any system, through its active usage, storage in backup systems, until final disposal. We are lead to believe that our private data are well-secured and protected against all kinds of adversaries who try to obtain such type of information.

Unfortunately, constantly surging security breaches show that all the it systems may be targeted by attackers. The organizations all over the world are no longer asking themselves *if* their IT systems are going to be under attack, but realized that the more appropriate question now is *when* their IT systems are going to be under attack. Therefore, system owners are undertaking countless of countermeasures to decrease the probability of performing a successful attack on their IT infrastructure, however, as the studies show, they still need to find a better solution of facing the insider threat.

The insider threat is one of the least remediated ones as its often underestimated in the risk analyses, yet globally it's one of the main sources of the IT security incidents. This paper focuses on the selected data security aspects from the insider threat perspective by drawing attention to one of the basic but uninvestigated enough attack vector – the inference attacks.

2. Data breaches

Along with the global digitalization of the data, a new threat landscape has emerged. The number of data breaches in the recent years is on the drastic rise and the volume of data leaked due to malicious acts grows significantly every year. The estimated global average cost of a data breach is 3.6 million dollars [1]. However, the real number and cost of such incidents remain unknown as the public statistics lack information about those breaches which are undisclosed to the media nor industry researchers.

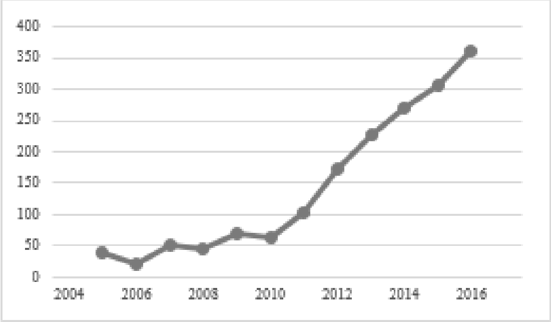

Fig. 1. The number of database records leaked between 2004 and 2016 in millions [10]

The article is based on the presentation awarded with the Polish General Inspector for Personal Data Protection Prize during the XXIII TI Forum.

Taken into account most publicly commented security breaches, one can draw a conclusion that, the majority of known data breaches is caused by external attackers who take over the control of a targeted system's database by exploiting some particular system's vulnerability [9]. However, digging deeper into analyses of the industrial data breaches reports [1], [2], [3], it becomes obvious that many organizations miss the mark on protecting the databases against insider threat, as the employees are indicated as the main source of the critical cyber security incidents.

The negligence or conscious actions of the employees despite being the leading cause of security incidents, remain also the least reported issue and can exist undetected for years. Therefore, organizations must undertake a different approach to protect the data from misuse by the employees than when securing it against external threats.

3. Security in database lifecycle

The database lifecycle [DBLC], which consists of five major stages, represents the phases through which the global schema of the database is planned, developed, evaluated, implemented and maintained in software-specific environments [7]. The DBLC stages divide into following:

1. *Requirements analysis* – information regarding the purpose and the natural data relations are gathered. The database related software is being selected.
2. *Logical design* – a conceptual data model is created with ER or UML techniques.
3. *Physical design* – access methods, partitioning and clustering of the data are assigned to increase the efficiency of the database.
4. *Implementation* – the formal schema is implemented using the data definition language.
5. *Maintenance* – existing database is monitored; the performance is analyzed and modification are continuously implemented. This stage lasts until any considerable modification is required to be made that leads to the database re-design or re-plan for another cycle of implementation.

Along with defining the database development as the continuous process, the security aspect became an imperative to be embedded into each stage of the lifecycle making the data and the database security management one of the inevitable elements to consider in this continuous process. The security planning begins at the early development stage with defining the data sensitivity levels and verifying legal aspects of data protection defined in the applicable laws.

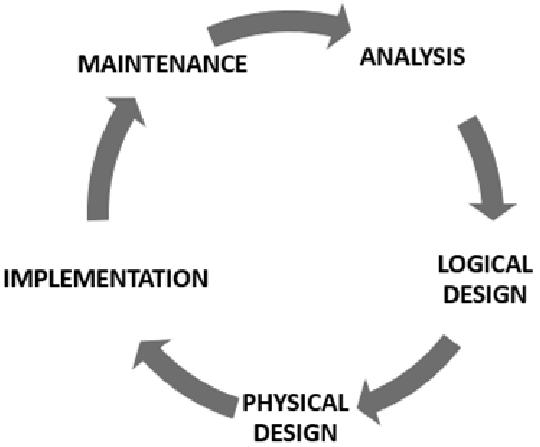

Fig. 2. Database lifecycle phases

From the design until the maintenance stage, the security considerations scope usually covers the software level preventive methods, e.g. physical and access controls, servers and software hardening, network communication security. What is very frequently missed is the database design level security. The data access for the organizations' employees is only controlled by a set of privileges. When not limited by law, the database owners do not enforce strict rules regarding the internal data access, giving unlimited competencies for the data processing and changes for some system accounts. Taken into account that the insider threat is currently the largest grossing attack vector for the IT systems, such approach cannot be considered satisfactory, even for the systems in which the data are not exposed to the public network and are only processed internally.

The data design level security is especially prone to security breaches in the following in the cases: when the excessive privileges are given to a selected group of access accounts over live system, the backup copies are improperly secured and maintained or some services (e.g. maintenance, accounting, etc.) are outsourced to third parties which require partial data access. The security at the data level should be a focal point of the design level, both logical and physical. If the appropriate design level limitations are later properly applied in the implementation phase, then majority of risks related with the data level vulnerabilities are mitigated. The following chapter describe selected protection methods for the data layer, widely used in the database systems to hide classified data from unauthorized access. However, despite the existing protection methods, there still exists a threat of inference which is later described in Chapter 5.

4. Data layer protection methods

When asked about data protection methods, the first obvious answer is encryption. This widely used technique, when applied properly, allows to effectively secure data, starting at record level up till the BLOB level. However, this method is only a subcategory of the much more comprehensive term which is pseudonymization. Along with anonymization, pseudonymization was one of the topic of discussion since the introduction of the European Union’s General Data Protection Regulation [GDPR]. The GDPR defines the understanding of those two terms and treats them as the privacy-enhancing factors of stored data. Although the context of the regulation is focused on handling personal data, both techniques may be applied universally when protecting the general data layer.

Tab. 1. Sample showing the difference between anonymization and pseudonymization

Original data	**Anonymized**	**Pseudonymized**
John Doe	XXXX	6cea57
Anna Smith	XXXX	739510
John Doe	XXXX	6cea57
Anna Doe	XXXX	d97fbf

Anonymization and pseudonymization are two distinct methods that permit to use de-identified data. The difference rests on whether the data can be re-identified. The common definition of anonymization, also mentioned in recital 26 of the GDPR, states that during the process *data rendered anonymous in such a way that the data subject is not or no longer identifiable* [4]. The data after the anonymization process must be impossible to recover even for the party who run this process, e.g. when anonymizing a selection of records containing personal data, the personal data may be replaced with random values and all the software logs and cache which potentially might still contain original data must be cleared. The data anonymization in theory is a straight-forward process however many organizations often fall short of actually anonymizing data, especially when the redundancy level of the stored data is high.

The pseudonymization is defined as the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information [4]. In this process the selected data is replaced by a different value, but unlike the anonymization, the data may be re-identified, by deriving the original data from the pseudonymized data using, as defined, “additional information”. In this category fall i.a. encryption, hash functions, tokenization. Again, the process itself is conceptually simple, however the proper execution and separation of the pseudonymized data from the information allowing to retrieve the original data remains a challenge.

For the mentioned insider threat scenarios, the pseudonymization seems to be more appropriate as the data may be temporarily concealed and recovered when necessary. However, there is still a threat of a security breach in case the data needed for recovery are captured along with the pseudonymized data, e.g. when using encryption as the pseudonymization method, the encryption keys are available under the same privileged account as the pseudonymized data.

5. Inference – the underestimated attack vector

The anonymization and pseudonymization should provide a certainty that the original data are impossible to re-identify. However, attackers when targeting the data layer, including the insider threat, may use more sophisticated class of attacks which is the inference.

The inference attacks derive sensitive information from non-sensitive information and the available metadata. The basic methods of inference attacks on statistical databases were described in [5], but the concept of majority of them is not limited to statistical databases and can be used in targeting all types of databases.

NON SENSITIVE DATA
METADATA
SENSITIVE DATA

Fig. 3. Inference attacks theory

One of the most known example of such attack was described in [6]. The supposedly de-identified data sets of GIC (Group Insurance Commission), which included medical records of individuals, were made public with assurance of the governor that the that GIC had protected patient privacy by deleting identifiers. Nonetheless, a researcher found a way to cross the available non-sensitive data with the publicly available metadata leading to full disclosure of sensitive data, including identification of governor's medical health records.

The same paper cites also the landmark study [8] which states that 87 percent of the U.S. population can be identified based on just three data points: five-digit ZIP code, gender, and date-of-birth. Analyzed separately those three data points are non-sensitive data as do not allow for any type of identity recovery. However, when analyzed together makes it possible to identify an individual.

Another more recent example was breaking anonymity of the Netflix customer data by cross-referencing the anonymized Netflix customer dataset with the public movie review information on IMDb [9].

The inference attacks are very frequently overlooked during the risk analyses and even when evaluated, the risk level related with them is estimated at negligible level. Nevertheless, in the light of the GDPR, the inference attacks should be revisited by data owners and controllers in occurrences of data which is said to be anonymized or pseudonymized data.

6. Solving inference threat

The inference attacks threat can be mitigated by adopting a mechanism called differential privacy. The differential privacy allows a data solicitor to collect data and infere meaningful information from the data without individual record attribution, i.e. the mechanism allows a solicitor to collect sensitive data, but the data cannot be attributed to any party.

The application of differential privacy involves several techniques, including the injection of mathematical noise in the collect data, data hashing, subsampling and randomized data injection [5]. Finding an appropriate tradeoff between the data accuracy and anonymity while adding a noise to the data remains a challenge as datasets' users need precise data to work on and the data owners need to comply with the data privacy regulations.

On top of the technical implementation of the differential privacy, the data security procedure must be added in function of the preventive, detective and corrective controls at each stage of DBLC. Without proper definitions of the data security standards and adhering to them, even the most relevant mathematical countermeasures may fail.

7. Conclusions

In the today's world all kinds of sensitive data are becoming a valuable currency, thus many intent to steal it. The race between the adversaries and security officers is gaining its momentum and it seems that nothing will ever stop it.

The data privacy is currently a popular research topic due to new EU legislation and organisations are now more determined than ever before to protect the data from external threats. What is frequently overlooked is an insider threat. Unlimited access to the data, lack of security-in-depth controls and growing demand for private data on the black market makes it tempting for the insiders to take advantage of security breaches. However, even more damaging from the conscious insider attackers is the negligence of both: the security departments and targeted victims. Capability to fully address all the possible attack vectors remains merely a wishful thinking, therefore continuous risk analysis is the pivotal point of the security of modern enterprises.

Along with personal information burst in the global network facilitated by social media, it seems that the privacy keeps losing its importance and people no longer appreciate this quality. However, when an individual's private data leaks without his consent into public, then the privacy unexpectedly becomes a substantial value. When processing personal data, the organizations must live up to expectations to keep disclosing personal data a choice of an individual, not becoming an unintended action. Therefore, risk analyses conducted by the security teams must analyze all the feasible attack vectors, including those very particular scenarios as inference attacks.

8. Bibliography

[1] *2016 Cost of Data Breach Study: Global Analysis. Ponemon Institute Research Report*, Ponemon Institute LLC, June 2016.

[2] *2016 Cost of Insider Threats Benchmark Study of Organizations in the United States*. *Ponemon Institute Research Report*, Ponemon Institute LLC, September 2016.

[3] *Fourth annual 2017 data breach industry forecast*, Experian Data Breach Resolution.

[4] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[5] Dzięgielewska O., Szafrański B., "A brief overview of basic inference attacks and protection controls for statistical databases", *Computer Science and Mathematical Modelling*, No. 4, 19–24 (2016).

[6] Ohm P., "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization", *UCLA Law Review*, Vol. 57, 1701–1777 (2010).

[7] Teorey T.J., Lightstone S.S., Nadeau T., Jagadish H.V., *Database Modeling and Design, Fifth Edition: Logical Design*, Morgan Kaufman, 5th Edition, Burlington, 2011.

[8] Sweeney L., *Simple Demographics Often Identify People Uniquely*, Carnegie Mellon University, Pittsburgh 2000.

[9] Narayanan A., Shmatikov V., *How to Break Anonymity of the Netflix Prize Dataset*, arXiv:cs/0610105v2 [cs.CR]

[10] Data breaches statistics available at: http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

Anonimizacja, tokenizacja, szyfrowanie. Jak odtworzyć nieodtwarzalne dane

O. DZIĘGIELEWSKA

Prywatność danych jest obecnie szeroko komentowana wśród wszystkich organizacji przetwarzających dane osobowe w związku z wprowadzeniem Rozporządzenia o Ochronie Danych Osobowych. Często zakłada się, że istniejące metody ochrony danych są wystarczające, ponieważ spełniają wymagania podejścia opartego na analizie ryzyka, w którym to koszt metod ochrony nie może przekroczyć wartości zasobu, w tym przypadku danych. Jednak liczba publicznie znanych przypadków wycieków danych potwierdza, że założenie to jest niepoprawne. Następstwa tego rodzaju incydentów bezpieczeństwa w niezliczonych przypadkach dowodzą, że podejście oparte na ryzyku nie spełniło swojej roli, ponieważ konsekwencje związane z utratą reputacji i stratami finansowymi znacznie przekraczają pierwotne szacowania. W artykule podkreśla się znaczenie ochrony warstwy danych od planowania, przez projektowanie, aż do etapów utrzymania w cyklu życia bazy danych, ponieważ liczne wektory ataku mające źródło wewnątrz organizacji i skierowane na warstwę danych wciąż przechodzą niezauważone podczas analizy ryzyka.

Słowa kluczowe: cykl życia bazy danych, ataki wnioskowaniem, prywatność danych, wyciek danych, RODO.