RISK MODEL OF APPLICATION OF LIFTING METHODS

A. Bogdanov[1], A. Degtyarev[1,3], N. Shchegoleva[1], V. Khvatov[4], N. Zaynalov[2], J. Kiyamov[1], A. Dik[1,a], A. Faradzhov[1]

[1] *Saint Petersburg State University, 7-9 Universitetskaya emb., Saint Petersburg, 199034, Russia*

[2]*Samarkand branch Tashkent university of information technology, Uzbekistan*

[3] *Plekhanov Russian University of Economics, 36 Stremyanny lane, Moscow, 117997, Russia*

[4]*DGT Technologies AG, http://dgt.world/*

E-mail: [a] st087383@student.spbu.ru

The article discusses the main provisions (methods, risk models, calculation algorithms, etc.) of the issue of organizing the protection of personal data (PD), based on the application of anonymization procedure. The authors reveal the relevance of the studied problem based on the tendency of the general growth of informatization and the further development of the Big Data technology. This circumstance leads to the need to use the so-called risk approach based on calculating the risk of PD as a probabilistic assessment of the amount of possible damage that the owner of the data resource may incur as a result of a successfully carried out information attack. For this purpose, the article describes an algorithm for calculating the risk of PD and proposes a risk model of the depersonalization procedure, which considers confidentiality problems arising both as a result of unauthorized access and as a consequence of planned data processing. To describe the risk model of the anonymization procedure, the types of attacks on the confidentiality of personal data, anonymization metrics and equivalence classes are analyzed, as well as the attacker's profiles and data distribution scenarios. Thus, the choice of a risk model for the depersonalization procedure was justified, and calculations for the generated synthetic set of PDs were presented. As a conclusion, it should be noted that the model of anonymization risk assessment proposed and tested on synthetic data makes it possible to abandon the concept of guaranteed anonymized data, introducing certain boundaries for working with risks and building a continuous process for assessing PD threats, taking into account the constantly growing volume of stored and processed information.

Keywords: information protection, personal data, depersonalization, information systems, model, risk of depersonalization procedure.

Alexander Bogdanov, Alexander Degtyarev, Nadezhda Shchegoleva, Valery Khvatov, Nodir Zaynalov, Jasur Kiyamov, Aleksandr Dik, Anar Faradzhov

1. Introduction

Recently, the problem of personal data protection (PD) has become more and more urgent. In this aspect, the question concerning the peculiarities of using various methods of depersonalization and related options for building a data risk model (risk model) is increasingly being raised. The relevance of this topic in the modern world is connected with the further introduction of information technologies into our lives, we are becoming more and more dependent on information systems and services, and, consequently, more and more vulnerable to security threats. Information systems that process personal data are particularly vulnerable to this risk. It is enough to remember the growth of unauthorized dissemination of personal data and its consequences in recent years. As examples, we should mention the theft of information about subscribers of mobile operators and other means of communication, trading information about bank customers, insurance companies, etc. In connection with these circumstances, it is advisable to consider the use of various methods of depersonalization as promising and potential ways to protect personal data. The process of depersonalization of data is part of the processing of personal data aimed at deleting identifying personal information, as a result of such a process, new depersonalized secure data is formed based on the initial array of information with personal data.

2. General description of the risk model

Currently, the international practice of using depersonalization methods is shifting towards a risk approach. In this case, the risk assessment is carried out in order to develop measures to ensure the confidentiality of private information if it is necessary to publish depersonalized data. The emergence of new sources of information makes it possible to compare data with previously published ones, which inevitably leads to the appearance of risks of re-identification. This, in turn, forces us to abandon the concept of guaranteed anonymized data, introducing certain boundaries of working with (risk threshold) and building a continuous process of assessing threats to personal data. As part of the standard approach, the risk is assessed based on the identification of threats (associated with the profile of the intruder) and existing vulnerabilities. At the same time, it should be taken into account that external and internal connections have a significant impact on the risk assessment: the availability of additional information, the motivation of the attacker, the legal framework, the IT systems used, management practices, etc. This leads to a division of risk between the risks of the data itself (taking into account the methods of depersonalization used) and the risks of the environment (contextual risks). Threats to the confidentiality of personal data arise as a result of authorized data processing, as well as as a result of unauthorized access or actions of an attacker.

3. Risk model

Taking into account the above factors, the study suggests considering the option of building a risk model based on the combined use of methods for assessing data risks and contextual risks. In order to carry out the risk assessment procedure, it is necessary to build a risk model that will determine the risk factors and the relationships between them, based on the following sequential steps:

- Risk factorization (identifying a set of individual risk components and establishing a link between them);

- Formation a release model of data;

- Setting quantitative risk thresholds;

- Determination of the necessary level of usefulness of the received depersonalized data;

- Justification of the procedure for constructing a risk model for a specific depersonalization procedure, including the possibility of re-evaluating the risk when using various depersonalization methods.

Conducting depersonalization taking into account the risk model requires a balance between the usefulness of the data obtained as a result of depersonalization in accordance with various estimated quantitative metrics (indicators) and an acceptable amount of risk. The risk thresholds are set in accordance with the use scenarios (public data, inter-organizational or private access). Within the framework of the model under consideration, depending on the purpose and objectives of depersonalization, the following quantitative metrics (indicators) will be used:

- risk level - the product of damage by the probability of the risk of re-identification;

- data utility level or data quality assessment;

- reversibility level, which allows you to maintain the connection of the original and depersonalized data set;

- variability of the depersonalization method.

- flexibility, which evaluates the possibility of making additions (distortions) to the array of depersonalized data.

- the resistance of an impersonal set to attacks is determined by the probability of success of re-identification attacks

- compatibility of various impersonal sets (when comparing attributes), etc.

In this case, the algorithm for calculating the risk of choosing a depersonalization strategy is as follows (Figure 1)

№	Step	Description
1.	Defining the goals of depersonalization of data	Explicit and documented goals of depersonalization of data, including the choice of a data distribution scenario
2.	Data structure analysis	Analysis of the structure of a specific set: selection of direct identifiers and quasi-identifiers based on the scenario and available targets, as well as likely attacks. Explicit replacement or exclusion of direct identifiers
3.	Determination of an acceptable level of risk (risk threshold)	Choosing a risk threshold based on the scenario and use case
4.	Contextual risk assessment	Conducting a scoring calculation for contextual risk
5.	Selecting parameters of the data risk calculation model	Analysis of depersonalization methods, model configuration
6.	Conducting depersonalization by a suitable method	Choosing a depersonalization method and building an impersonal data set U (depersonalized)
7.	Data risk assessment	Conducting a data risk assessment taking into account the size of the equivalence class and the chosen anonymization measure (taking into account the choice of the attacker's profile) - for data sets D and U (original and depersonalized)
8.	Calculating the total risk	Risk calculation, damage assessment
9.	Evaluating the usefulness of data	Calculation of data utility metrics and making a decision about using this model or repeating the process 1-8 with other parameters
10.	General risk assessment	Risk reduction and comparison with existing restrictions on the level of risk. If necessary, the construction of additional metrics and the repetition of calculations
11.	Completion of the procedure	Preparing a publication or refusing to depersonalize, documenting the procedure

Figure 1. Algorithm for implementing the risk model

In this algorithm, contextual risks and data risks are calculated separately.

4. Features of building data risks

As for contextual risks, they are an assessment of categorizable factors of organizational and technical impact on the organization of the process of storing and converting personal data. Taking into account the impact of these factors and their mutual influence, it is proposed to implement an

assessment of contextual risks on the basis of a scoring model, based on the implementation of the risk calculator [1], while modeling the score card is carried out on the basis of the linear regression method.

Data risks are understood as the risks of re-identification associated with the structure and composition of data. Access to such data may be obtained as a result of errors on the part of third parties, service personnel or through applications (for example. REST API). In this case, it is advisable to include the following methodological operations in the composition of the recommended model.

1. Processing attributes to highlight direct identifiers and quasi-identifiers

The risk of using depersonalized data consists in identifying a specific individual in the data set and assigning to it those attributes that are contained in the set. This situation is called re-identification. From the point of view of assessing the risk of re-identification, the most important are sensitive attributes that, in the case of compromise, disclosure or illegal use, can lead to significant damage, embarrassment and/or inconvenience. According to [3], it is customary to distinguish:

- Direct identifiers (used directly)

- Quasi-identifiers (used in combination)

2. Planning possible re-identification attacks

Under attacks on the confidentiality of PD, we will understand unauthorized actions on the part of an attacker aimed at re-identifying the records of an individual inside an impersonal data set [2]. The data risk assessment applied to a specific set of depersonalized data depends on the depersonalization methods chosen - for suppression or aggregation. The selection of appropriate quasi-identifiers requires taking into account various types of attacks, which can be combined:

- re-identification attacks through linkage, linkage attacks is an attempt to identify an individual through linking two sets of data;

- attribution attack is carried out through the disclosure of attributes: the transfer to an individual of the attributes of the group to which he supposedly belongs;

- subtraction attack is aimed at reducing the original data set at the expense of additional knowledge;

- Inference attack - collecting available information to attack a more secure system;

- differentiation attack involves the identification of a person's personality on the basis of additional information about him, allowing us to assume his dissimilarity to the majority;

- reconstruction attack is aimed at existing sets of aggregated data.

3. Definition of anonymization level metrics.

As a result of using depersonalization methods, data with varying degrees of anonymization is obtained. There are several measures to measure anonymization. Most of them are based on the concept of an equivalence class – the ability to allocate identical records within a data set in terms of quasi-identifiers. Anonymity metrics are closely related to the frequency analysis of records, the probability of re-identification is generally inversely proportional to such metrics.

There are the following types of metrics and attacks on them:

- k-anonymity;

- ℓ-diversity;

- t-closeness;

4. Determination of utility level metrics.

For a large amount of data, it is important to have quantitative estimates of the usefulness of data that show the quality of data after applying depersonalization methods. Utility metrics (data quality) can be quite complex. In this regard, it is recommended to use more than one of the following:

- General metrics of information loss

- Classification metric

- Reuse metrics

- Entropy-based information loss metric

- A measure of mutual utility;

5. Selecting the attacker's profile

When calculating the risk probabilities, it is important to take into account the types of attacks that can be generalized into attacker profiles. It is assumed that the attacker has the necessary resources and knowledge to carry out the necessary attacks. The goals and availability of access to additional information vary [4]. In accordance with the established tradition, the name of the profiles is compared with three groups: "Marketer", "Prosecutor" and "Journalist".

6. Identification of scenarios for the distribution of depersonalized data

Scenarios for the distribution of Release models play an important role in the process of depersonalization, since they require various degrees of depersonalization. For example, for the public dissemination of data, a higher level of protection is required. Data distribution scenarios depend on several decisions that affect contextual risk and data risk, as shown in Figure 2 [5]:

Sections	The script "Public Data"	The script "Inter-organizational interaction"	The script "Private data"
Access rights	Everyone has free access to the published data	Access to the published data (or part of it) is provided to a limited number of persons or organizations	A narrow circle of individuals or organizations has access to the published data
Data usage	Unlimited access to data through the web portal, that is, free access for everyone	• Ensuring security on its territory • Access granted • Remote virtual access • Access via the analytical server	Exchange within and between organizations
entitlement	Unlimited rights to multiple use and dissemination of data	Available from an authorized person or organization	The re-use, replication and dissemination of data is prohibited
Re-identification attempt	Demo attack or use as additional data	• Deliberate internal attack • Unintentional identification of a specific person in the data set by a friend • Data leak	

Figure 2. Data distribution scenarios

5. Conclusion

It should be noted that the proposed combined version of the risk assessment model makes it possible to comprehensively (at the level of contextual risks and data risks) conduct a detailed analysis, and then a balanced choice of the method of depersonalization of personal data necessary for application both at the enterprise and on a national scale. As a result, this circumstance brings novelty and prospects to the solution of the issue under consideration.

References

[1] Handbook on Security of Personal Data Processing, ENISA, 2017.

[2] The Anonymization Decision-Making Framework, UKAN, 2016.

[3] General Data Protection Regulation (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).

[4] Goldberger, J. and T. Tassa (2010). Efficient anonymizations with enhanced utility. Transactions on Data Privacy 3 (2), 149–175.

[5] Framework of de-identification process for telecommunication service providers, ITU-T X.1148, 2020.