Preserving Privacy in Personal Data Processing

Cansu Saatci
Dept. of Computer Engineering
Eskisehir Osmangazi University
Eskisehir, Turkiye
cansu.saatci@gmail.com

Efnan Sora Gunal
Dept. of Computer Engineering
Eskisehir Osmangazi University
Eskisehir, Turkiye
esora@ogu.edu.tr

***Abstract*—Nowadays, the spread of information systems and the operation of service sectors through digital applications such as data mining, data analysis, recommendation systems, and digital marketing have brought the need to store and process large amounts of personal data in the systems of organizations. Although it is beneficial to process these data within the scope of the services provided, it may cause victimization and loss of rights if unauthorized persons take over. Therefore, the protection methods to be applied to these personal data have become a necessity for all organizations. Ensuring privacy during the processing, storage and publishing (sharing) of data is an important concern. In recent years, privacy measures have gained more importance due to national and international regulations for the protection of personal data as well. Consequently, there are certain rules that organizations should follow in their information systems, where large volumes of personal data are processed, stored and shared. In this work, some of the techniques, which are used for preserving privacy while processing personal data, are reviewed.**

Keywords— data privacy, general data protection regulation, law on the protection of personal data, personal data

I. INTRODUCTION

Privacy is the right of individuals to decide when, by whom and to what extent their information is used. All data collecting and processing institutions are responsible for the safe storage of this data.

The concepts of data privacy and data protection are highly interrelated, so they are often confused with each other. Actually, data protection refers to securing data against unauthorized access, while data privacy implies that under what circumstances and by whom data can be processed, stored and transmitted. Data protection is basically a technical issue and the presence of relevant information can be protected by the necessary protection procedures. On the other hand, ensuring the privacy of the data is a necessity at the constitutional level [1].

A. Personal Data

Personal data can be defined as any information that is suitable for identifying individuals. In this context, the person's identity, communication, health and financial information as well as personal life, religious belief and political opinion are considered as personal data [2].

The constitutional court does not only disclose personal data such as name, surname, date of birth and place of birth of the individual; but also the data that makes the person directly or indirectly identifiable such as telephone number, motor vehicle license plate, social security number, passport number, resume, picture, image and sound recordings, fingerprints, genetic information, IP address, e-mail address, hobbies, preferences, interacting persons, group memberships, and family information. These personal data are frequently used in our daily lives [3].

B. Sensitive Personal Data

Sensitive personal data consists of health, sexual life, religion, sect and other beliefs, race, ethnicity, philosophical belief, political thought, associations, foundations, trade union membership, clothing choices, criminal convictions and security measures, biometric and genetic data. If these data are seized by unauthorized people, the owners of the data may be exposed to victimization or discrimination. Therefore, further protection is needed for sensitive personal data [2].

C. Law No. 6698 on Protection of Personal Data (LPPD)

The purpose of the Law on the Protection of Personal Data is to discipline the processing of personal data and to protect fundamental rights and freedoms, the privacy of private life. Protection of the right of privacy and information security of the people is also considered within this scope. In addition, it is among the objectives of the Law to regulate the obligations and procedures and principles to be complied with by real and legal persons that process personal data [4].

Law on the protection of personal data is mainly prepared to prevent unlawful processing of personal data, to prevent unlawful access to personal data and to provide safe storage of personal data [2].

D. General Data Protection Regulation - GDPR

The European Union (EU) General Data Protection Regulation (GDPR) was developed to draft a series of harmonized data privacy laws to protect EU citizens throughout Europe. GDPR came into effect on 25 May 2018. It regulates the use of all kinds of personal data for all companies that process the personal data of anyone living within the European Union, including the ways in which companies collect, share and use data, regardless of where the company is located. All organizations that monitor their behavior are affected by GDPR. The regulations apply to both controllers and handlers. Also, All Turkish companies that process the personal data of EU citizens are also obliged to comply with GDPR [5].

GDPR and LPPD have severe sanctions for data breaches, but the aim of both regulations is to provide safe protection of personal data.

E. Privacy Preserving Data Publishing (PPDP)

Privacy Preserving Data Publishing [6] provides some methods and tools for publishing data while protecting privacy. Each record has many attributes. In PPDP, attributes are analyzed in 4 categories. These categories are explained below:

i) Personal Information Identifiers (PII): Information that can directly identify an individual's identity. Name, ID or other information that is linked to an individual, and so on.

978-1-7281-3992-0/19/$31.00 ©2019 IEEE

ii) Quasi-Identifiers: Information that can identify an individual's identity when combined with a feature (or set of attributes).

iii) Sensitive Attributes: Personal data that is not expected to be disclosed.

iv) Non-Sensitive Attributes: Information that is neither confidential nor personal and can be published.

In the following sections, privacy in digital marketing, recommender systems, and health research is discussed. Then, some of the common techniques for preserving data privacy are explained and compared. Finally, conclusions of the work are presented.

II. Privacy in Digital Marketing and Recommender Systems

Digital marketing is the marketing of products or services that uses digital technologies, especially through the internet, but also in smart, display advertising and other digital environments.

With the spread of digitalization, users are demanding online shopping sites and online services. This makes it difficult for consumers to choose from thousands of products and services. Companies have started to use product recommendation systems in order to compete in such a big market. Recommender systems speed up the decision-making process of the users. They analyze the behaviors, profiles and likes of users and try to present other products and product groups that they may be interested in. This is a very useful method for marketing activities. However, in order to carry out these activities; the data generated during the interaction of users with digital systems is shared with third parties [7]. The sharing may cause victimization and loss of rights if the necessary precautions are not taken. Therefore, preserving privacy is crucial during the design and use of recommender systems [8].

III. Privacy in Health Research

The protection of personal data in health studies is especially important in terms of patient privacy and is the obligation of the physician to keep it secret [9]. The obligation to keep secrets is a sub-obligation arising from the loyalty obligation of the health workers. Also, health care providers and health researchers have an obligation to maintain the confidentiality of patient information too. Therefore, necessary precautions must be taken during the storage, processing and sharing of patient information.

IV. Techniques for Preserving Data Privacy

In the following subsections, some of the common techniques to preserve data privacy are explained. However, it should be noted that not all of these methods may be appropriate for the analysis needs of the data being studied on. Different techniques might be needed as well in certain cases.

A. Anonymization

The anonymization of personal data means that the data cannot be associated with any particular or identifiable person, even by pairing it with other data. There is no legal drawback in processing and sharing anonymized data. It is widely used in statistical research.

B. k - Anonymity

Sweeny proposed a model called k-anonymity in her study to ensure privacy [10]. In this model, k-anonymity is used as a feature of the dataset to define the anonymity level of the dataset. In general, it means that there is at least *1/k* possibility to correctly associate with a person by looking at the properties in a data row. In this case, the larger the number, the stronger the anonymity of the dataset.

As an example, a dataset shown in Table I contains the age, gender and test results of the students in a class. This dataset is called as k-2 k-anonymous because each age-gender pair has at least two rows: (13, M), (14, F), (15, M). That is, even if it is known that there are two 15-year-old male students in the classroom, this registration cannot be determined to be exactly 5th or 6th. There are ½ possibilities for correct detection.

TABLE I. K-2 ANONYMITY

No	Age	Gender	Test Result
1	13	M	83
2	13	M	67
3	14	F	89
4	14	F	76
5	15	M	82
6	15	M	91

However, when the dataset given in Table II is examined, this dataset has a k-1 anonymity since there is only one 15-year-old male or female student so that the combination of (15, F) and (15, M) describes certain individuals. Person identification can be made with 100% probability in that case [11].

TABLE II. K-1 ANONYMITY

No	Age	Gender	Test Result
1	13	M	83
2	13	M	67
3	14	F	89
4	14	F	76
5	15	M	82
6	15	F	91

When the k-anonymity method is used, it is ensured that the record of another analysis (or information) is not associated with a person. This is called identity disclosure problem in literature. However, there are two types of attacks that the k-anonymity method cannot solve. These are homogeneity and background knowledge attacks. What these types of attacks are examined through the medical record data of an imaginary hospital as follows:

Table III shows the medical records of an imaginary hospital [12]. The table does not contain unique descriptive attributes such as name and social security number. In this example, the attributes are divided into two groups. Sensitive attributes (consisting solely of medical condition) and non-sensitive attributes (postal code, age and nationality). If an attribute is marked as sensitive, the value of that attribute of any individual in the dataset must not be discovered. Also, the combination of {postal code, age, nationality} attributes is the quasi-identifier of this dataset.

Table IV shows the 4-anonymous form derived from Table III, where '*' symbol refers to a suppressed value, e.g. postal code = '1485*' means that the postal code is in the

range of (14850−14859). Similarly, age = '4*' means that age is between 40 and 49. In Table IV, each group has the same values as the semi-identifier with at least three other records in the table.

TABLE III. MEDICAL RECORDS

No	Non-Sensitive			Sensitive
	Postal Code	Age	Nationality	Condition
1	13054	37	Canadian	Heart Attack
2	13068	35	American	Heart Attack
3	13068	32	Japanese	Headache
4	13054	36	American	Headache
5	14853	62	Indian	Diabetes
6	14853	65	Canadian	Heart Attack
7	14850	54	American	Headache
8	14850	59	German	Headache
9	13054	41	American	Diabetes
10	13054	48	Indian	Diabetes
11	13068	46	Japanese	Diabetes
12	13068	43	American	Diabetes

TABLE IV. 4-ANONYMOUS MEDICAL RECORDS

No	Non-Sensitive			Sensitive
	Postal Code	Age	Nationality	Condition
1	130**	< 40	*	Heart Attack
2	130**	< 40	*	Heart Attack
3	130**	< 40	*	Headache
4	130**	< 40	*	Headache
5	1485*	≥ 50	*	Diabetes
6	1485*	≥ 50	*	Heart Attack
7	1485*	≥ 50	*	Headache
8	1485*	≥ 50	*	Headache
9	130**	4*	*	Diabetes
10	130**	4*	*	Diabetes
11	130**	4*	*	Diabetes
12	130**	4*	*	Diabetes

Because of its conceptual simplicity, k-anonymity is popular and widely preferred in the literature for methods of securing privacy. However, anonymity does not always guarantee privacy. In the following subsection, attacks to k-anonymity are briefly discussed.

Attacks of k-anonymity

The k-anonymity method is vulnerable to two types of attacks, namely homogeneity attacks and background knowledge attacks based on experience. These attacks are used for data disclosure. In the following, these two attack types are described with example scenarios:

In the first scenario, Bella and Tom are two neighbors. One day, Tom gets sick and is taken to the hospital by ambulance. Seeing the ambulance, Bella begins to discover what illness Tom is suffering from. Bella has the 4-anonymous table of current inpatient records published in the imaginary hospital (Table IV) and knows that one of the records in this table contains Tom's data. Since Bella is Tom's neighbor, she knows that Tom is a 41-years-old American man and he lives in postal code 13054. Therefore, Bella knows that Tom's patient registration number is 9, 10, 11 or 12. Patients in this group have the same medical condition (diabetes), and so Bella can conclude that Tom is diabetes. This is referred to as Homogeneity Attack. k-anonymity may cause information leakage due to lack of diversity [12].

In the second scenario, Bella has a friend, Rieko, who was admitted to the same hospital as Tom and whose patient records appear in Table IV. Bella knows that Rieko is a 32-years-old Japanese woman who currently lives in the postal code 13068. Based on this information, Bella can learn that Rieko's information is in records 1, 2, 3 or 4. Without additional information, Bella is not sure if Rieko has a headache or if she has a heart attack. However, it is known that the Japanese have a very low rate of heart attack. Using this information, Bella may decide that Rieko has a headache. Hence, this is an example of Background Knowledge Attack. In this case, k-anonymity does not protect against background-based attacks.

Since both attacks are plausible in real life, a stronger definition of privacy is needed, considering diversity and knowledge based on experience. Subsequent studies have proposed the *l*-diversity method, which is protected against these vulnerabilities.

TABLE V. 3-DIVERSE MEDICAL RECORDS

No	Non-Sensitive			Sensitive
	Postal Code	Age	Nationality	Condition
1	1305*	≤ 50	*	Heart Attack
4	1305*	≤ 50	*	Headache
9	1305*	≤ 50	*	Diabetes
10	1305*	≤ 50	*	Diabetes
5	1485*	> 50	*	Diabetes
6	1485*	> 50	*	Heart Attack
7	1485*	> 50	*	Headache
8	1485*	> 50	*	Headache
2	1306*	≤ 50	*	Heart Attack
3	1306*	≤ 50	*	Headache
11	1306*	≤ 50	*	Diabetes
12	1306*	≤ 50	*	Diabetes

C. *l*-Diversity

If the values for one or more hidden properties in all records are the same, the values for those properties for target T can be learned. The purpose of *l*-diversity is to contain at least one instance of the relevant sensitive data category in each stored and organized data group. Since each group contains 3 different sensitive data in Table V, the dataset can be called 3-diversity [12].

If Bella and Tom sample is examined on 3-diverse dataset; Since Bella is Tom's neighbor, she knows that Tom is a 41-years-old American man and he lives in postal code 13054. However, there are 3 different sensitive data in the 4 records included in this group. Therefore, Bella cannot determine which of these four records belongs to Tom. In this case, the desired diversity is achieved.

D. Masking

It is the deletion or masking of certain areas of personal data, making the person indeterminable. For example, a part of a person's credit card number is masked with asterisk so that it looks like "4866 **** **** 0002". There may be a need to access to personal data in many business roles, but it should not be exploited using this access authority. Also, many organizations may need to use production, or live, data for their application developers to perform extensive testing. In such cases, data masking may be needed when sharing data. Many database management systems have data masking capabilities.

E. Derivation

The detailed data is replaced with the more general ones. For example, the day / month / year details of the date of birth

are replaced directly with the age of the person. In this way, anonymization is performed by deriving data.

F. Mixing

It means the destruction of the detectability of individuals without damaging the total benefit by mixing the values in the dataset. In a class where the average age is to be taken, the values showing the age of the individuals are changed with each other so that the data is mixed.

G. Approximation

Approximation approach is a technique used to replace certain personal data with less specific values. For example, the birth date of a user (August 20, 1995) is stored in the form of July 1 - September 25, 1995, or even only 1995 [13].

H. Encryption

Encryption is a process that encrypts data into an unreadable format (cipher text) so that only people or systems accessing the appropriate key can decrypt and read it. Encryption can protect personal information, emails and other sensitive data as well as secure network connections. Organizations are increasingly use encryption to protect applications and sensitive information from reputational damage. Although it is not a compulsory technique in GDPR, it is mentioned as a proposal [13].

I. Tokenization

Tokenization is to replace a piece of data with a unique identifier that acts as a proxy that can be used to retrieve the original value. Tokenization is often used in online payment solutions and systems where secure API connections are required [13].

J. Adding Noise

Adding noise is the process of making random, statistically insignificant changes to a dataset that maintains the processing capability when identifying personal data. The addition of noise can greatly alter the data and reduce the reliability of the data. Therefore, it is not a technique that is used very often [13].

V. Comparison of Techniques

The techniques used to ensure privacy in data processing vary according to the structure, size and purpose of the data processed by organizations. For a single problem, it is often not possible to apply all the techniques mentioned in section IV together. Even a technique that is suitable for some datasets can produce very unsuccessful results for another dataset. It is therefore difficult to make a full comparison of these techniques. However, the new dataset obtained by the application of these techniques can be compared based on three factors, namely data reliability, the amount of data being modified, and the level of anonymity provided, as given in Table VI.

VI. Conclusions

While processing personal data (either sensitive or not), preserving the privacy is of greater importance. In recent years, national and international regulations also define the measures to follow while processing personal data. Since data processing is required to improve services in almost all fields (i.e., health, digital marketing and so on), researchers are constantly working on new approaches to preserve privacy while processing the data.

TABLE VI. Comparison of Techniques

Techniques	Data Reliability	Data Modification	Data Anonymity
Anonymization	***	**	***
k-Anonymity	***	-	**
l-Diversity	***	-	***
Masking	***	-	**
Derivation	**	**	**
Mixing	**	***	**
Approximation	**	**	**
Encryption	***	-	***
Tokenization	***	-	***
Adding Noise	*	***	*
(-: None, *: Low, **: Medium, ***: High)			

References

[1] R. Robinson, “Data privacy vs. data protection,” 2018. [Online]. Available: https://blog.ipswitch.com/data-privacy-vs-data-protection. [Accessed: 05-Agu-2019].

[2] TBMM, Kişisel Verilerin Korunması Kanunu - KVKK. Turkey, 2016.

[3] Anayasa Mahkemesi E. 2013/122 K. 2014/74 kararı.

[4] TBMM, “Kişisel Verilerin Korunması Kanunu Tasarısı (1/541) ve Adalet Komisyonu Raporu Sıra Sayısı: 117.” [Online]. Available: https://www.tbmm.gov.tr/sirasayi/donem26/yil01/ss117.pdf. [Accessed: 05-Agu-2019].

[5] Microsoft, “GDPR.” [Online]. Available: https://products.office.com/tr-tr/business/articles/5-things-to-know-about-gdpr-before-its-too-late. [Accessed: 05-Agu-2019].

[6] J. Vasa and P. Modi, “Review of different privacy preserving techniques in PPDP,” Int. J. Eng. Trends Technol., vol. 59, no. 5, pp. 223–227, 2018.

[7] M. Yüksektepe, “Kişisel verilerin korunması ve dijital pazarlama,” 2018. [Online]. Available: https://www.webtures.com.tr/blog/kisisel-verilerin-korunmasi-ve-dijital-pazarlama/. [Accessed: 05-Agu-2019].

[8] Cerebro, “Ürün öneri sistemleri ile pazarlama,” 2018. [Online]. Available: https://medium.com/@cerebro.tech/ürün-öneri-sistemleri-ile-pazarlama-e07e9d358e5f. [Accessed: 05-Agu-2019].

[9] M. V. Dülger, “Sağlık hukukunda kişisel verilerin korunması ve hasta mahremiyeti,” İstanbul Medipol Üniversitesi Hukuk Fakültesi Dergisi, vol. 1, no. 2, pp. 43–80, 2015.

[10] L. Sweeney, “k-anonymity: a model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557–570, 2002.

[11] O. Angiuli, “What is k-anonymity.” [Online]. Available: https://www.quora.com/What-is-k-anonymity. [Accessed: 05-Agu-2019].

[12] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “ℓ-Diversity: Privacy beyond k-anonymity,” In Proc. of IEEE 22nd International Conference on Data Engineering, vol. 2006, p. 24, 2006.

[13] N. Farrell, “Techniques for personal data privacy,” 2017. [Online]. Available: https://www.cuttlesoft.com/techniques-for-personal-data-privacy/. [Accessed: 05-Agu-2019]