Chapter 2
Existing Privacy Protection Solutions

In this chapter, we outline the major developments of modern privacy study based on the survey work we have conducted [1–4]. Mainstream privacy protection techniques including anonymity, clustering-based, differential privacy, cryptography, and machine learning methods will be presented in the following sections.

2.1 Preliminary of Privacy Study

In this section, we present an overview of privacy systems, including different participation roles, anonymization operations, and data status. We also introduce the terms and definitions of the system.

In terms of participants, we can see four different roles in the privacy protection domain.

- Data generator: Individuals or organizations who generate the original raw data (e.g., medical records of patients, bank transactions of customers), and offer the data to others in a way either actively (e.g. posting photos to social networks to the public) or passively (leaving records of credit card transactions in commercial systems).
- Data curator: The people or organizations who collect, store, hold, and release the data. Of course, the released data sets are usually anonymized before publishing.
- Data user: The people who access the released data sets for various purposes.
- Data attacker: The people who try to gain more information from the released data sets with a benign or malicious purpose. We can see that a data attacker is a special kind of data user.

There are three major data operations in privacy-preserving systems.

- Collecting: Data curators collect data from different data sources.

Y. Qu et al., *Personalized Privacy Protection in Big Data*, Data Analytics,
https://doi.org/10.1007/978-981-16-3750-6_2

Table 2.1 A table of patients in a medical database

Name	Job	Gender	Age	Disease	Other
Linda	Singer	F	30	FLU	NA
Allen	Researcher	M	25	Fever	NA
...	...	...	...	...	...

- Anonymizing: Data curators anonymize the collected data sets in order to release it to public.
- Communicating: Data users performan information retrieval on the released data sets.

Furthermore, a data set of the system possesses one of the following three different statuses.

- Raw: The original format of data.
- Collected: The data has been received and processed (such as de-noising, transforming), and stored in the storage space of the data curators.
- Anonymized. The data has been processed by an anonymization operation.

We can see that an attacker could achieve his goals by attacking any of the roles and operations. In general, we can divide a given record into four categories according to its attributes.

- Explicit identifier: A unique attribute that can clearly identify an individual, such as passport ID and drive licence numbers.
- Quasi-identifier: Attributes that be used to identify individuals with a high probability by combining other information, such as gender, birthday, age, etc. In fact, different attackers will have different quasi-identifiers according to their background knowledge.
- Sensitive attributes: The expected information interested by an adversary. In general, it is difficult to predict in advance.
- Non-sensitive attributes: The information not in the previous three categories.

We provide an example as shown in Table 2.1. In the example, *name* is an explicit identifier, while *work*, *gender*, and *age* constitute a set of quasi-identifiers, *disease* is sensitive information.

2.2 Anonymity Based and Clustering Based Methods

The data clustering direction developed from the initial k-anonymity method, then the l-diversity method, and then the t-closeness. We use Table 2.1 as an example to quickly demonstrate the journey of the data clustering methods for privacy protection.

Table 2.2 An illustration of k-anonymity (k = 2)

Job	Gender	Age	Disease	Other
Artist	M	20–30	FLU	NA
Artist	M	20–30	HIV	NA
Professional	F	30–40	FLU	NA
Professional	F	30–40	Cancer	NA

In 2000, the strategy of k-anonymity protection is shown in Table 2.2, with the strategy that each record in the table is at least as identical to the other $k-1$ records on the quasi-identifier. Thereby reducing the probability of being identifiable. As shown in the example, dancers, singers, etc. are merged into artists, lawyers, and engineers are combined into the professional occupation, and the accurate age is expressed as a range. The k value in this example is 2. In this way, the maximum probability that a patient can be identified is $\frac{1}{k}$. If the k value is large enough, patient privacy can be effectively protected. It can be mathematically described as follows.

Let $T = t_1, t_2, ..., t_n$ be a table of a data set D, $A = A_1, A_2, ..., A_m$ be all the attributes of T, and $t_i[A_j]$ be the value of attribute A_j of tuple t_i. If $C = C_1, C_2, ..., C_k \subseteq A$, then we denote $T[C] = t[C_1], t[C_2], ..., t[C_k]$ as the projection of t onto the attributes in C.

The quasi-identifier is defined as a set of non-sensitive attributes of a table if these attributes can be linked with external data sets to uniquely identify at least one individual in the data set D. We use QI to represent the set of all quasi-identifiers.

A table T satisfies k-anonymity if for every tuple $t \in T$ there exist at least $k-1$ other tuples $t_{i_1}, t_{i_2}, ..., t_{i_{k-1}} \in T$, such that $t[C] = t_{i_1}[C] = t_{i_2}[C], ..., t_{i_{k-1}}[C]$, for all $C \in QI$.

On the other hand, we can also note that a larger k value will result in more data loss. At the same time, under the homogenous attack, the k-anonymity model cannot effectively protect the privacy of users due to the homogeneity of sensitive attributes. For example, the attacker knew that Linda was in Table 2.2 and she had cancer. Based on this background knowledge, the attacker knows that Linda is the fourth record in the table.

To overcome the shortcomings of the k-anonymity model, Machanavajjhala et al. [48] proposed the l-diversity model in 2006, requiring at least one sensitive attribute value is different in each anonymous group. In this way, the probability of an attacker can infer a certain record of private information is up to $\frac{1}{l}$. Table 2.3 provides a concrete example, where $k = 2, l = 2$.

As aforementioned, l-diversity [23] is an extension of the k-anonymity to "well represent" the sensitive attributes. In particular, there are four different interpretations of the term "well represented" as follows.

(1) Distinct l-diversity. Similar to k-anonymity, each sensitive attribute has to possess at least l distinct values in each qid group.

Table 2.3 An Illustration of l-anonymity (k = 2, l = 2)

Job	Gender	Age	Disease	Other
Artist	F	20–30	HIV	NA
Artist	F	20–30	HIV	NA
Artist	F	20–30	Cancer	NA
Artist	F	20–30	Cancer	NA

(2) Probabilistic l-diversity. The frequency of a sensitive value in a qid group is at most $\frac{1}{l}$.

(3) Entropy l-diversity. For every qid group, its entropy is at least $\log l$.

(4) (c, l)-diversity. The frequency of sensitive values of a qid group is confined in the range defined by c (a real number) and l (in integer).

However, the l-diversity based method cannot prevent the similarity attack, as the attacker can infer the sensitive information of the user according to the sensitive familiarity value and the semantic similarity of each QI-group. In some specific scenarios, the l-diversity model may provide more background knowledge for attackers.

In order to solve the above problems, Li et al. proposed t-Closeness in 2010. The specific strategy is: for a given QI-group, ensure that the difference between its distribution and the corresponding distribution on the original data set does not exceed a certain threshold. Based on the above three models, researchers further developed some protection methods, such as (a, k)—anonymous [5], (k, e)—nonymous [6], and (e, m)—Anonymous [7], etc. However, the anonymity-based protection models require special attack assumptions, and cannot perform quantitative analysis. Therefore, it has great limitations in practical applications.

2.3 Differential Privacy Methods

Different from the data clustering strategy, the differential privacy framework [25] was proposed in 2006, which offers strong privacy protection in sense of information theory. The basic background is that an attacker may obtain expected information by multiple queries to a statistical database on top of his background knowledge of victims. The defense strategy is: for two data sets with a minimum difference, the difference between the queries on the two data sets is very limited, therefore limiting the information gain for attackers. One popular method to achieve this is adding noise to outputs.

Definition 2.1 *Differential Privacy:* A random function M satisfies ϵ-differential privacy if for every $D_1 \sim D_2$, and for all outputs $t \in P$ of this randomized function, the following statement holds:

$$P_r[M(D_1)] \leq exp(\epsilon) P_r[M(D_2)], \tag{2.1}$$

in which exp refers to the exponential function. Two data sets D_1 and D_2 are neighbours with at most one different item. ϵ is the privacy protection parameter that controls the degree of difference induced by two neighbouring data sets. A smaller ϵ leads to a stronger privacy guarantee.

We can achieve $\epsilon-$differential privacy by adding random noise whose magnitude is adjusted according to the global sensitivity.

Definition 2.2 *Global Sensitivity:* The global sensitivity $S(f)$ of a function f is the maximum absolute difference obtained on the output over all neighbouring data sets:

$$S(f) = \max_{D_1 \sim D_2} |f(D_1 - D_2)|. \tag{2.2}$$

Two mechanisms are always utilized to satisfy the differential privacy definition: The Laplace mechanism and the Exponential mechanism. Between these two mechanisms, the Laplace mechanism achieves $\epsilon-$differential privacy by adding noise that following Laplace distribution is more suitable for numeric outputs.

Definition 2.3 *Laplace Mechanism:* Given a function $f: D \rightarrow P$, the mechanism $M{:}R \rightarrow \triangle(R)^n$ adds Laplace distributed noise to the output of f:

$$M(D) = f(D) + V, where V \sim Lap\Big(\frac{S(f)}{\epsilon}\Big), \tag{2.3}$$

where $Lap\Big(\frac{S(f)}{\epsilon}\Big)$ has PDF $\frac{1}{2\sigma} exp(\frac{-\epsilon|x|}{\sigma}), \sigma = \frac{S(f)}{\epsilon}$ is the scale parameter. The novel algorithm developed in this paper adopts the standard Laplacian mechanism.

Lee and Clifton [28] found that differential privacy does not match the legal definition of privacy, which is required to protect individually identifiable data, rather than the how much one individual can affect an output as differential privacy provides. As a result, they proposed differential identifiability to provide strong privacy guarantees of differential privacy, while letting policy-makers set parameters based on the established privacy concept of individual identifiability. Following this research line, Li et al. [29] analyzed the pros and cons of differential privacy and differential identifiability and proposed a framework called membership privacy. The proposed framework offers a principled approach to developing new privacy notions under which better utility can be achieved than what is possible under differential privacy.

As differential privacy is a global concept for all users of a given data set, namely the privacy protection granularity is the same to all protected users, therefore it is called uniform privacy or homogenous differential privacy. In order to offer customized privacy protection for individuals, personalized differential privacy (also named as heterogenous differential privacy or non-uniform privacy) was also extensively explored [30, 42].

2.4 Cryptography Based Methods

Based on the current situations in practice, we can conclude that encryption is still the dominant methodology for privacy protection.

Cryptography can certainly be used in numerous fashions for privacy protection in the big data age. For example, a patient can use the public key of her doctor to encrypt her medical documents and deposits the ciphertext into the doctor's online database for her treatment while her privacy is strictly preserved.

With the emergence of big data, clouds are built to serve many applications due to its economical nature and accessibility feature. For example, many medical data sets are outsourced to clouds, which triggers privacy concerns from patients. The medical records of a patient can only be accessed by authorized persons, such as her doctors, rather than other doctors or people. The public key encryption is obviously not convenient if the number of authorized people is sufficiently large due to the key management issue. In this case, Attribute-Based Encryption (ABE) is an appropriate tool [8, 9], which was invented in 2004 by Sahai and Waters [10]. In the ABE scheme, a set of descriptive attributes of the related parties, such as hospital ID and doctor ID are used to generate a secret key to encrypt messages. The decryption of a ciphertext is possible only if the set of attributes of the user key matches the attributes of the ciphertext. The ABE scheme creatively integrates encryption and access control, and therefore no key exchange problem among the members of the authorized group.

The dilemma of encryption-based privacy protection in big data is: on one hand, we need to offer sufficient privacy protection for users, at the same time, we have to make the encrypted data informative and meaningful for big data analysis and public usage. As a result, we face a number of challenges as follows. One challenge is information retrieval on encrypted data. This research branch is also called searchable encryption, which boomed around the year 2000 [11, 12]. The basic idea is as follows. An user indexes and encrypts her document collection, and sends the secure index together with the encrypted data to a server that may be malicious. To search for a given keyword, the user generates and submits a trapdoor for the keyword, which the server uses to run the search operation and recover pointers to the appropriate encrypted documents.

Another challenge here is operations on encrypted objects. This research branch is named as homomorphic encryption started in 1978 [48]. In this kind of encryptions, we expect to carry out computations on ciphertext, and obtain an encrypted output. If we decrypt the output it should match the result of operations performed on the original plaintext. Mathematically, we can describe it as follows: given a message m, a key k, and an encryption algorithm E, we can obtain a ciphertext $E_k(m)$. Let f be a function, and its corresponding function is f', D_k be a decryption algorithm under key k, then an encryption scheme is homomorphic if $f(m) = D_k(f'(E_k(m)))$.

In 2009, Gentry kicked off a further development in this direction, Fully Homomorphic Encryption (FHE), which supports arbitrary computation on ciphertexts [13]. A survey on this branch can be found in [14]. The problem is that we do not have a feasible fully homomorphic encryption system in place yet due to the

extraordinary inefficiency in computing. Compared to FHE, Multi-Party Computation (MPC), which was initiated by Yao in 1982 [15], has been used in practice by offering weaker security guarantees but much more efficient. The scenario of MPC is like this: multiple participants jointly compute a public function based on their private inputs while reserving their input privacy against the other participants, respectively.

We have to note that encryption can protect the privacy of an object itself, however, it is vulnerable against side information attacks, such as traffic analysis attacks against anonymous communication systems. For example, we can encrypt web pages of a protected website, however, the encryption cannot change the fingerprints of the web pages, which are represented by the size of the HTML text, number of webobjects, and the size of the web objects. An attacker can figure out which web pages or web sites a victim visited using traffic analysis methodology [16–18]. In terms of solutions, information theory based packet padding is the main player, including dummy packet padding [19] and predicted packet padding [20].

2.5 Machine Learning and AI Methods

The flourishing of machine learning (ML) has become one of the drivers of privacy concerns in modern society. Sensitive information of users may be compromised during the data collecting and model training process. Fortunately, recent studies have shown that some ML methods can also act as tools for privacy protection if employed correctly. Novel decentralized learning framework which can facilitate distributed learning tasks and enable source data to remain on edge devices has received widespread attention [21].

Distributed training system contains the following main modules: data and model partitioning module, stand-alone optimization module, a communication module, as well as data and model aggregation module. In particular, different machines are responsible for different parts of the model and assigned with different data. Therefore, distributed training systems can keep datasets containing privacy at different locations instead of the cloud, which have been widely applied in recent years.

Edge computing is a widely applied decentralized architecture that performs processing tasks in intelligent edge nodes. Similar to distributed training, the architecture of edge computing can mitigate privacy issues [22]. However, other security techniques are required to combine with. For example, Gai et al. [23] combined blockchain and edge computing techniques to address the security and privacy issues in smart grid. Ma et al. [24] proposed a lightweight privacy-preserving classification framework for face recognition by employing additive secret sharing and edge computing. Li et al. [25] proposed a privacy protection data aggregation scheme for mobile edge computing assisted IoT applications based on the Boneh-Goh-Nissim cryptosystem. Du et al. [26] handled with privacy problems of training datasets and proposed a differential privacy based protection method in wireless big data with edge computing.

Federated learning (FL), as a novel distributed learning paradigm, becomes prominent recently to address privacy issues. Different from traditional methods which put all data $D_1 \cup D_2 \cup ... \cup D_N$ to train a model M_{SUM}, a FL system is a learning process in which the data owners collaboratively train a model M_{FED}, while any data owner F_i dose not share her data D_i to others [27]. In addition, let V_{FED} represent the accuracy of M_{FED}, it should be very close to V_{SUM}. Specifically, let δ denote a non-negative real number, if

$$|M_{FED} - M_{SUM}| < \delta, \tag{2.4}$$

we say that the federated learning algorithm has $\delta-$accuracy loss.

However, recent studies have demonstrated that the framework of FL also has some privacy issues [28]. One major concern is that adversaries could recover sensitive data by violating the shared parameters. To mitigate this problem, Bonawitz et al. [29] designed a secure aggregation method to protect the privacy of each user's model gradient. Recently, Liu et al. [30] pointed out that user dropout and untrusted server are two unresolved challenges of original FL schemes. Thus, they proposed a robust federated extreme gradient boosting framework for mobile crowdsensing that supports forced aggregation. Hao et al. [31] proposed a privacy-enhanced FL scheme by employing homomorphic ciphertext and differential privacy. The proposed noninteractive method can achieve both effective protection and efficiency. From the perspective of verifying whether the cloud server is operating correctly, Xu et al. [28] proposed a privacy-preserving and verifiable FL framework based on double-masking protocol.

References

1. J. Yu, K. Wang, D. Zeng, C. Zhu, S. Guo, Privacy-preserving data aggregation computing in cyber-physical social systems. ACM Trans. Cyber-Phys. Syst. **3**(1), 1–23 (2018)
2. L. Cui, G. Xie, Y. Qu, L. Gao, Y. Yang, Security and privacy in smart cities: challenges and opportunities. IEEE Access **6**, 134–146 (2018)
3. Y. Qu, M.R. Nosouhi, L. Cui, S. Yu, Privacy preservation in smart cities, in *Smart Cities Cybersecurity and Privacy* (Elsevier, Amsterdam, 2019), pp. 75–88
4. Y. Qu, S. Yu, W. Zhou, S. Peng, G. Wang, K. Xiao, Privacy of things: emerging challenges and opportunities in wireless internet of things. IEEE Wirel. Commun. **25**(6), 91–97 (2018)
5. R.C.-W. Wong, J. Li, A. W.-C. Fu, K. Wang, (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2006), pp. 754–759
6. Q. Zhang, N. Koudas, D. Srivastava, T. Yu, Aggregate query answering on anonymized tables, in *IEEE 23rd International Conference on Data Engineering* (IEEE, 2007), pp. 116–125
7. J. Li, Y. Tao, X. Xiao, Preservation of proximity privacy in publishing numerical sensitive data, in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (ACM, 2008), pp. 473–486
8. V. Goyal, O. Pandey, A. Sahai, B. Waters, Attribute-based encryption for fine-grained access control of encrypted data, in *Proceedings of the 13th ACM Conference on Computer and Communications Security* (2006), pp. 89–98

9. A. Lewko, B. Waters, Decentralizing attribute-based encryption, in *Annual International Conference on the Theory and Applications of Cryptographic Techniques* (Springer, 2011), pp. 568–588
10. A. Sahai, B. Waters, Fuzzy identity-based encryption, in *Annual International Conference on the Theory and Applications of Cryptographic Techniques* (Springer, 2005), pp. 457–473
11. D.X. Song, D. Wagner, A. Perrig, Practical techniques for searches on encrypted data, in *Proceeding of the IEEE Symposium on Security and Privacy. S&P 2000* (IEEE, 2000), pp. 44–55
12. R. Curtmola, J. Garay, S. Kamara, R. Ostrovsky, Searchable symmetric encryption: improved definitions and efficient constructions. J. Comput. Secur. **19**(5), 895–934 (2011)
13. C. Gentry, Fully homomorphic encryption using ideal lattices, in *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing* (2009), pp. 169–178
14. V. Vaikuntanathan, Computing blindfolded: new developments in fully homomorphic encryption, in *IEEE 52nd Annual Symposium on Foundations of Computer Science* (IEEE, 2011), pp. 5–16
15. A.C. Yao, Protocols for secure computations, in *23rd Annual Symposium on Foundations of Computer Science (sfcs)* (IEEE, 1982), pp. 160–164
16. Q. Sun, D.R. Simon, Y.-M. Wang, W. Russell, V.N. Padmanabhan, L. Qiu, Statistical identification of encrypted web browsing traffic, in *Proceedings of the IEEE Symposium on Security and Privacy* (IEEE, 2002), pp. 19–30
17. M. Liberatore and B. N. Levine, Inferring the source of encrypted http connections, in *Proceedings of the 13th ACM Conference on Computer and Communications Security* (2006), pp. 255–263
18. Y. Zhu, X. Fu, B. Graham, R. Bettati, W. Zhao, Correlation-based traffic analysis attacks on anonymity networks. IEEE Trans. Parallel Distrib. Syst. **21**(7), 954–967 (2009)
19. P. Venkitasubramaniam, T. He, L. Tong, Anonymous networking amidst eavesdroppers. IEEE Trans. Inf. Theory **54**(6), 2770–2784 (2008)
20. S. Yu, G. Zhao, W. Dou, S. James, Predicted packet padding for anonymous web browsing against traffic analysis attacks. IEEE Trans. Inf. Forensics Secur. **7**(4), 1381–1393 (2012)
21. Y. Sun, J. Liu, J. Wang, Y. Cao, N. Kato, When machine learning meets privacy in 6g: a survey. IEEE Commun. Surv. Tutorials (2020)
22. J. Zhang, B. Chen, Y. Zhao, X. Cheng, F. Hu, Data security and privacy-preserving in edge computing paradigm: survey and open issues. IEEE Access **6**, 209–237 (2018)
23. K. Gai, Y. Wu, L. Zhu, L. Xu, Y. Zhang, Permissioned blockchain and edge computing empowered privacy-preserving smart grid networks. IEEE Internet Things J. **6**(5), 7992–8004 (2019)
24. Z. Ma, Y. Liu, X. Liu, J. Ma, K. Ren, Lightweight privacy-preserving ensemble classification for face recognition. IEEE Internet Things J. **6**(3), 5778–5790 (2019)
25. X. Li, S. Liu, F. Wu, S. Kumari, J.J. Rodrigues, Privacy preserving data aggregation scheme for mobile edge computing assisted iot applications. IEEE Internet Things J. **6**(3), 4755–4763 (2018)
26. M. Du, K. Wang, Z. Xia, Y. Zhang, Differential privacy preserving of training model in wireless big data with edge computing. IEEE Trans. Big Data (2018)
27. Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: concept and applications. ACM Trans. Intell. Syst. Technol. (TIST) **10**(2), 1–19 (2019)
28. G. Xu, H. Li, S. Liu, K. Yang, X. Lin, Verifynet: secure and verifiable federated learning. IEEE Trans. Inf. Forensics Secur. **15**, 911–926 (2019)
29. K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H.B. McMahan, S. Patel, D. Ramage, A. Segal, K. Seth, Practical secure aggregation for federated learning on user-held data (2016), arXiv preprint arXiv:1611.04482
30. Y. Liu, Z. Ma, X. Liu, S. Ma, S. Nepal, R. Deng, Boosting privately: privacy-preserving federated extreme boosting for mobile crowdsensing (2019), arXiv preprint arXiv:1907.10218
31. M. Hao, H. Li, X. Luo, G. Xu, H. Yang, S. Liu, Efficient and privacy-enhanced federated learning for industrial artificial intelligence. IEEE Trans. Industr. Inf. **16**(10), 6532–6542 (2019)