Mathematical Problems of Computer Science 55, 35–43, 2021.
doi: 10.51408/1963-0071

UDC 519.72

The Role of Information Theory in the Field of Big Data Privacy

Mariam E. Haroutunian[1] and Karen A. Mastoyan[2]

[1]Institute for Informatics and Automation Problems of NAS RA
[2]Gavar State University
e-mail: armar@sci.am, kmastoyan@yandex.com

Abstract

Protecting privacy in Big Data is a rapidly growing research area. The first approach towards privacy assurance was the anonymity method. However, recent research indicated that simply anonymized data sets can be easily attacked. Later, differential privacy was proposed, which proved to be the most promising approach. The trade-off between privacy and the usefulness of published data, as well as other problems, such as the availability of metrics to compare different ways of achieving anonymity, are in the realm of Information Theory. Although a number of review articles are available in literature, the information - theoretic methods capacities haven't been paid due attention. In the current article an overview of state-of-the-art methods from Information Theory to ensure privacy are provided.
Keywords: Big data, Anonymization, Differential privacy, Entropy, Mutual information, Distortion.
Article info: Received 20 February 2021; accepted 18 April 2021.

1. Introduction

In recent years, Big Data has become a hot research topic, because it helps businesses and organizations to improve the decision making power and provides new opportunities with data analysis.

Big Data life cycle can be divided into the following stages: data generation, storage and processing. Multiple parties are involved in these stages, hence, the privacy violation risks are increased.

A number of privacy preserving mechanisms have been developed [1], [2], however, the study on Big Data privacy issues are at a very early stage [3]. Modern technologies and tools, such as social networks, search engines, hacking packages, data mining and machine learning tools, cause a lot of problems to individual privacy.

In general, it is very hard to find a clear definition or a global measurement on privacy. The studies on privacy can be separated into two classes: content privacy and interaction

privacy. So far, the majority of research on privacy protection is conducted in the context of databases. The goal of a privacy preserving statistical database is to enable the user to learn properties of the population while securing the personal information.

The practically dominant privacy protection strategy is the use of cryptography. One way to protect data is to encrypt it in such a way that only the owner can decrypt it. The task of machine learning is to find the dependency in the data. An idea proceeds: why not to train the model on encrypted data? The problem with this approach is that when we encrypt data, the dependencies in it are lost, because this is the aim of encryption - to change the data so that the dependencies cannot be discovered. The degree of entropy in the data after encryption prevents models from capturing these dependencies. Therefore, encrypting data and training models on them do not work. The other disadvantage of cryptography for privacy is the limited computing power of mobile devices for safe encryption and decryption algorithms. That is why other methods for privacy preserving are required.

The main research categories of privacy are the data clustering and the theoretical frameworks. **Anonymization** is a key component of data clustering. Anonymization is the process of removing personal identifiers, both direct and indirect, that can lead to a person's identification. A person can be directly identified by name, address, zip code, telephone number, photograph or image, or other unique personal characteristics. A person can be indirectly identified if certain information is linked to other sources of information, including workplace, job title, salary, zip code, or even the fact of having a specific diagnosis or condition. Data cannot be completely anonymous and useful. Generally speaking, the richer the data, the more interesting and useful it is. This has led to the concepts of anonymization and removal of personally identifiable information, by which it is hoped that sensitive parts of the data can be suppressed to maintain the confidentiality of the records, while the rest can be published and used for analysis.

The early approach in data clustering direction is the k-anonymity method (1998), then its extension as l-diversity was suggested in 2007 and later the t-closeness method was developed in 2010. In the second category of privacy frameworks the differential privacy (DP) and its developments are included (Fig. 1). DP neutralizes linkage attacks, since it is a property of the data access mechanism and is not related to the presence or absence of auxiliary information available to the attacker.

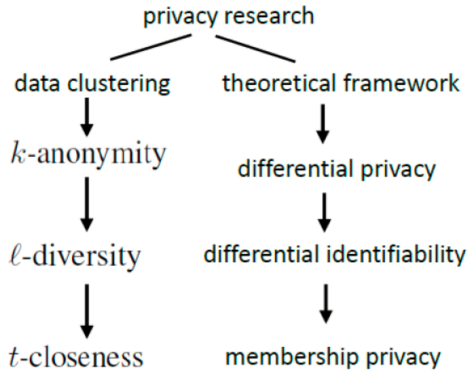

Fig. 1. The categories of privacy study [3].

When comparing syntactic and DP approaches, one can recall the trade-off between privacy and data utility. Finally, databases are passed on to provide certain benefits (for

example, research knowledge, such as the effectiveness of a medical procedure). On the one hand, in order to maximize the usefulness of the data, all data can be published untouched, thus completely breaching privacy. On the other hand, not publishing any of the data will lead to maximum privacy, but this empty database will be useless. Therefore, the organization should seriously consider both its interests in data exchange and the risks they are willing to accept. An organization seeking to exchange sensitive data should consider the logistics issues involved in the exchange process and carefully balance the confidentiality and usefulness of the published data. An important step in one of these considerations is the choice of syntactic and DP. Such problems can be solved by information - theoretic methods and tools.

Although, there are some survey/review - type papers introducing the concept of privacy protection in Big Data, none of them provide detailed discussion regarding privacy with respect to Information Theory. In the current paper we introduce a summary of up-to-date methods on privacy assurance from Information Theory standpoints.

The rest of the paper is structured as follows. The DP is discussed in section 2. The developments of privacy based on information - theoretic methods are presented in section 3. The paper is summarized in section 4.

2. Differential Privacy

The DP framework was suggested in 2006 [4], that offers privacy protection in the sense of Information Theory. In recent years this topic has attracted attention and has been researched in literature. Main results, among many others, are surveyed in [5] - [8].

DP examines impossibility paradox of obtaining any information about a specific person by studying useful information about a multitude of people. Suppose the trusted party contains a set of sensitive personal data (eg email usage data, movie watching data, medical records) and wants to provide global statistical information about it. Such a system is called a statistical database. By providing such aggregated statistical information about the data, it is possible to disclose some information about individuals.

DP ensures that data about individuals from such a database cannot be retrieved, no matter what additional datasets or sources of information are available to the attacker. Such a guarantee is achieved due to the fact that the owner of the database uses such a mechanism (algorithm) for providing data, in which the presence or absence of information about a person in the database will not significantly affect the result of the request to it.

DP is designed to maximize the accuracy of queries from statistical databases while minimizing the possibility of disclosing the anonymity of records. The problem of analyzing sensitive data has a long history spanning many areas. As data about people become more and more detailed and technology allows more and more of this data to be collected and analyzed, there is an increasing need for a reliable, mathematically rigorous definition of privacy, as well as a class of algorithms that satisfy this definition. Various approaches to anonymizing data have failed when researchers have been able to identify personal information by combining two or more separate statistical databases. DP is the basis for the formalization of confidentiality in statistical databases and was introduced in order to protect against such methods of disclosing anonymity (deanonymization).

DP allows users to protect and maintain privacy when their data is in a specific database, just as they would be safe, if the data were not in some database. After the publication of human data in the database, in accordance with the differentiated confidentiality, the

likelihood of violation of the confidentiality of people should not increase. That is, the degree of secrecy can be assessed by the likelihood of damage. This is one of the practical definitions of privacy.

DP can provide extremely strong guarantees of user privacy, but it does not guarantee unconditional relief from all damages. And it doesn't provide privacy where it didn't exist before. In general, DP does not guarantee that what a person considers his secret will remain secret. It simply ensures that participation in the survey is not disclosed by itself, that participation does not reveal any of the characteristics included in the survey. It is possible that the results of the survey may reflect statistical data about a person. Health screening for early signs of illness can produce strong, even convincing results. The fact that these findings are valid for humans does not imply a breach of confidentiality. The person may not even participate in the survey (DP ensures that these results are equally likely, regardless of whether the person participated in the survey or not).

It is desirable that DP be endowed with the following qualities: protection against arbitrary risks, automatic neutralization of linkage attacks, quantification of privacy loss.

DP is based on introducing randomness into data. To realize this, there are different mechanisms, e.g. the laplace mechanism, the exponential mechanism, mechanisms via α -nets, etc. Due to the fact that differential privacy is a probabilistic concept, any of its methods necessarily has a random component. Some of them, like Laplace's method, use the addition of controlled noise to the function to be calculated. Laplace's method adds Laplace noise, i.e. the noise from the Laplace distribution.

DP works by adding statistical noise to data (or its inputs or outputs). Depending on the location of the noise, DP is classified into two types: local DP and global DP (Fig. 2).

The most commonly used threat model in differential privacy is the **global DP model**. The main component is a trusted data curator. Each source sends him his confidential data, and it collects them in one place (for example, on a server). A repository is trusted if we assume that it processes our sensitive data on its own, does not transfer it to anyone, and cannot be compromised by anyone. In other words, we believe that a server with sensitive data cannot be hacked. Within the central model, we usually add noise to query responses. The advantage of this model is the ability to add the lowest possible noise value, thus maintaining the maximum accuracy allowed by the principles of DP. The disadvantage of the central model is that it requires a trusted store, and many of them are not. In fact, the lack of trust in the consumer of the data is usually the main reason for using DP principles.

The **local DP model** allows you to get rid of the trusted data store: each data source (or data owner) adds noise to their data before transferring it to the store. This means that the storage will never contain sensitive information, implying there is no need for its power of attorney. The local model of DP avoids the main problem of the central model: if the data warehouse is compromised, then hackers will only have access to noisy data that already meets the requirements of DP.

The local model is less accurate than the central one. In the local model, each source independently adds noise to satisfy its own differential privacy conditions, so that the total noise from all participants is much greater than the noise in the central model. Ultimately, this approach is only justified for queries with a very persistent trend (signal). Apple, for example, uses a local model to estimate the popularity of emoji, but the result is only useful for the most popular emoji (where the trend is most pronounced). Typically, this model is not used for more complex queries, such as those used by the US Census Bureau or machine learning. The central and local models have both advantages and disadvantages, and now

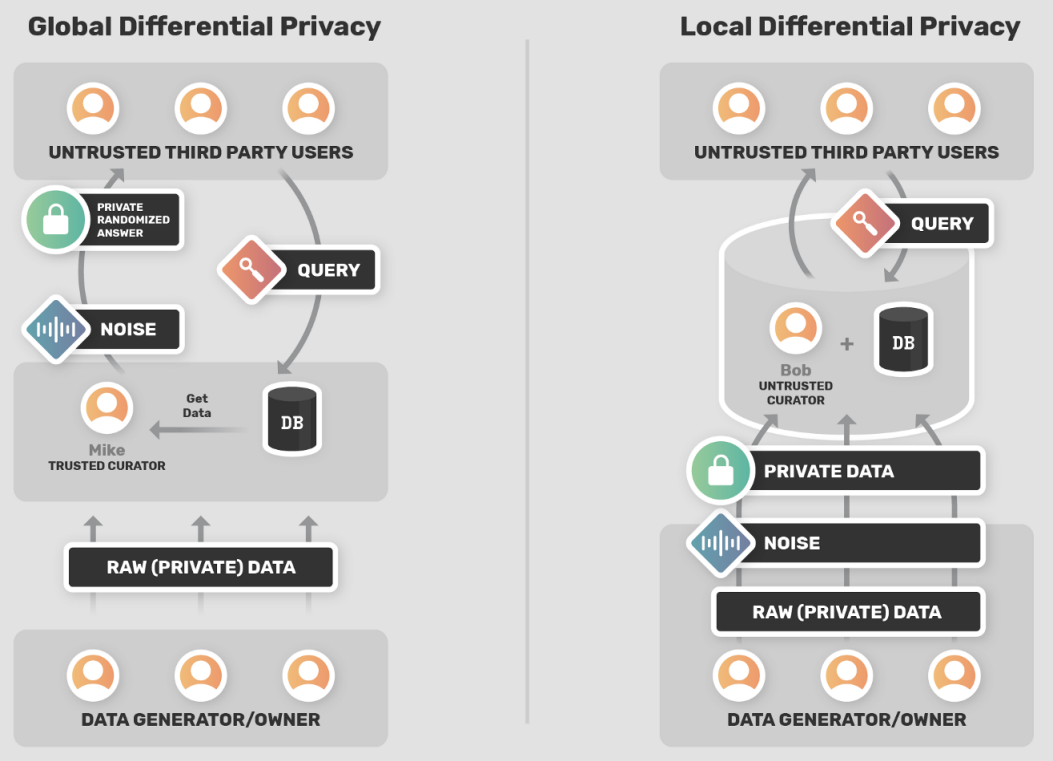

Fig. 2 Global Differential Privacy and Local Differential Privacy

the main effort is to get the best of them.

3. Review of Information Theory - based Results in Privacy

The first problem is to have a metric to compare various ways of achieving anonymity. The initial focus on analyzing the anonymity of messaging through mixed - based anonymity systems in which all network communication is available for the attacker is given in [9]. An information - theoretic metric based on the idea of anonymity probability distributions is introduced. In the same paper it is demonstrated that if maximum route length in the mix system exists, it is known to the attacker and can be used to extract additional information. It means that more advanced probabilistic metrics of anonymity are needed.

An analytical measure of anonymity of routs in eavesdropped networks is proposed in [10] using the information-theoretic equivocation. Cryptographic techniques prevent analysis of packet content, however, information can be gained by analyzing the correlation of transmission schedules of multiple nodes, as the packet timing information is easy to obtain in wireless networks. For anonymity it is necessary for the routes to be undetectable using the correlation across the transmission schedules, which results in a tradeoff between anonymity and network performance. For this purpose a quantifiable metric is defined in [10] using the uncertainty in networking information. The key result shows the equivalence between anonymity - performance tradeoff and information - theoretic rate distortion.

It is often important to allow researchers to analyze data without compromising the privacy of individuals or leaking confidential information outside the organization. In [11] it is shown that sparse regression for high dimensional data can be carried out directly on a compressed form of the data, in a manner to guard privacy in information - theoretic sense.

The suggested compression reduces the number of data records exponentially preserving the number of input variables. These compressed data can then be made available for statistical analyses with the same accuracy as the original data. In this case the original data are not recoverable from the compressed data, and the algorithms run faster, requiring fewer resources. The privacy (the problem of recovering the uncompressed data from the compressed one) is evaluated in information - theoretic terms by bounding the average mutual information, which is connected with the problem of computing the channel capacity of certain systems.

A new privacy measure in terms of Information Theory, similar to t-closeness is defined in [12], which can be achieved by the postrandomization method in the descrete case and by noise addition in the general case. The privacy criterion here is an average measure over a divergence, and the privacy - distortion problem is strongly related to the rate - distortion problem in the field of Information Theory, namely, the problem of lossy compression of source data subject to a distortion criterion.

A new measure for privacy of votes is proposed in [13], that relies on the notion of entropy. Entropy is a natural choice to measure privacy in an information - theoretic setting, and authors demonstrate how different formulations of conditional entropy answer different questions about vote privacy. A theorem has been established that enables accurate analysis of privacy offered by complex cryptographic voting protocols. Connections between two existing privacy notions for votes have been established.

The study of DP from a rate - distortion perspective has been initiated in [14]. Rate - distortion is applicable when the goal of the data collector is to publish an approximation of the data itself. The case when the data collector is not trusted is considered, which leads to using the local DP as a privacy measure. A robust rate-distortion setting is considered, in which the source distribution is unknown, but comes from some class. The goal is to look for a locally differentially private channel, that achieves minimum privacy risks while guaranteeing distortion of the given level.

In [15] the relation between three different notions of privacy: identifiability, differential privacy and mutual - information privacy is investigated. Identifiability guarantees indistinguishability between probabilities, DP guarantees limited additional disclosures, and mutual information is the information - theoretic notion. Under a unified privacy - distortion framework, where the distortion is the Hamming distance between the input and output databases, some connections between these three privacy notions have been established.

Guaranteeing a tight bound on privacy risk often incurs a significant penalty in terms of the usefulness of the published result. This privacy-utility tradeoff is studied in [16] in the context of publishing a differentially private approximation of the full data set and measure utility via a distortion measure.

4. Conclusion

In this article a general outlook on the current methods for estimating privacy of databases from Information Theory perspectives is provided. A series of publications devoted to various problems of privacy solved by information - theoretic tools and methods is analyzed. Research has shown that information-theoretic methods are effective for a wide range of tasks ranging from anonymity to differential privacy.

References

[1] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua and S. Guo, "Protection of Big Data Privacy", *IEEE Access*, vol. 4, pp. 1821–1834, 2016, doi: 10.1109/ACCESS.2016.2558446.

[2] L. Xu, C. Jiang, J. Wang, J. Yuan and Y. Ren, "Information security in Big Data: Privacy and Data Mining" *IEEE Access*, vol. 2, pp. 1149–1176, 2014, doi: 10.1109/ACCESS.2014.2362522.

[3] S. Yu, "Big privacy: Challenges and opportunities of privacy study in the age of Big Data", *IEEE Acces*, vol. 4, pp. 2751–2763, 2016, doi: 10.1109/ACCESS.2016.2577036.

[4] C. Dwork, M. Bugliesi, B. Preneel, V. Sassone, I. Wegener (eds) Automata, "Differential Privacy", *Languages and Programming. ICALP*, Lecture Notes in Computer Science, vol 4052, Springer, Berlin, Heidelberg, 2006. https://doi.org/10.1007/11787006 1

[5] K. M. P. Shrivastva, M. A. Rizvi and S. Singh, "Big Data privacy based on differential privacy a hope for Big Data," *Proc. Intern. Conf. on Computational Intelligence and Communication Networks*, Bhopal, India, pp. 776–781, 2014. doi: 10.1109/CICN.2014.167.

[6] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy", Foundations and Trends in Theoretical Computer Science: vol. 9, no. 3-4, pp 211–407. 2014. http://dx.doi.org/10.1561/0400000042

[7] N. Li, M. Lyu, D. Su and W. Yang, *Differential Privacy: From Theory to Practice*, Morgan & Claypool, 2016. doi: 10.2200/S00735ED1V01Y201609SPT018.

[8] X. Yao, X. Zhou and J. Ma, "Differential Privacy of Big Data: An Overview," *IEEE 2nd Intern. Conf. on Big Data Security on Cloud (BigDataSecurity), IEEE Intern. Conf. on High Performance and Smart Computing (HPSC), and IEEE Intern. Conf. on Intelligent Data and Security (IDS)*, New York, NY, USA, pp. 7–12, 2016. doi: 10.1109/BigDataSecurity-HPSC-IDS.2016.9.

[9] A. Serjantov and G. Danezis, "Towards an Information Theoretic Metric for Anonymity". *In: Dingledine R., Syverson P. (eds) Privacy Enhancing Technologies*, Lecture Notes in Computer Science, vol 2482. Springer, Berlin, Heidelberg, 2003. https://doi.org/10.1007/3-540-36467-6-4

[10] P. Venkitasubramaniam, T. He and L. Tong, "Anonymous networking amidst eavesdroppers," *IEEE Trans. on Information Theory*, vol. 54, no. 6, pp. 2770–2784, June 2008. doi: 10.1109/TIT.2008.921660.

[11] S. Zhou, J. Lafferty and L. Wasserman, "Compressed and privacy-sensitive sparse regression," *IEEE Trans. on Information Theory*, vol. 55, no. 2, pp. 846–866, Feb. 2009. doi: 10.1109/TIT.2008.2009605.

[12] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer, "From t-closeness-like privacy to postrandomization via Information Theory", *IEEE Trans. on Knowl. and Data Eng.*, vol. 22, no. 11, pp. 1623-1636, 2010. DOI:https://doi.org/10.1109/TKDE.2009.190

[13] D. Bernhard, V. Cortier, O. Pereira, and B. Warinschi, "Measuring vote privacy, revisited", *Proc. of ACM conf. on Computer and Communications Security*, Association for Computing Machinery, New York, NY, USA, pp. 941952, 2012. DOI:https://doi.org/10.1145/2382196.2382295

[14] A. Sarwate and L. Sankar, "A rate-distortion perspective on local differential privacy", *52 annual Allerton conf., UIUC*, Illinois, USA, pp. 903–908, 2014.

[15] W. Wang, L. Ying and J. Zhang, "On the relation between identifiability, differential privacy, and mutual-information privacy," *IEEE Trans. on Information Theory*, vol. 62, no. 9, pp. 5018–5029, Sept. 2016. doi: 10.1109/TIT.2016.2584610.

[16] K. Kalantari, L. Sankar and A. D. Sarwate, "Optimal differential privacy mechanisms under Hamming distortion for structured source classes," *IEEE Intern. Symp. on Information Theory*, Barcelona, Spain, pp. 2069–2073, 2016. doi: 10.1109/ISIT.2016.7541663.

Ինֆորմացիայի տեսության դերը մեծ տվյալների գաղտնիության ոլորտում

Մարիամ Ե. Հարությունյան[1] և Կարեն Ա. Մաստոյան[2]

[1]ՀՀ ԳԱԱ Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտ
[2]Գավառի պետական համալսարան
e-mail: earmar@sci.am, kmastoyan@yandex.com

Ամփոփում

Մեծ տվյալների հետ աշխատելիս գաղտնիության պահպանումը հետազոտության արագ աճող ոլորտ է: Գաղտնիության առաջին մոտեցումը անանունության մեթոդն էր: Վերջին ուսումնասիրությունները ցույց են տվել, որ պարզապես անանուն տվյալների շտեմարանները կարող են հեշտությամբ ենթարկվել հարձակման գաղտնիության տեսանկյունից: Հետագայում առաջարկվեց դիֆերենցիալ գաղտնիությունը, որն ապացուցեց, որ ամենահեռանկարայինն է: Գաղտնիության և հրապարակված տվյալների օգտակարության միջև փոխզիջումը, ինչպես նաև այլ հարցերի, ինչպիսիք են անանունության հասնելու տարբեր եղանակները համեմատելու համար չափի առկայությունը, ընկնում է ինֆորմացիայի տեսության տիրույթում: Չնայած գրականության մեջ մի շարք ակնարկային հոդվածների առկայությանը, ինֆորմացիայի տեսության մեթոդների հնարավորությունները պատշաճ ուշադրության չեն արժանացել: Այս հոդվածում մենք ներկայացնում ենք ինֆորմացիայի տեսության ժամանակակից մեթոդների ակնարկ՝ գաղտնիությունն ապահովելու համար: Վերլուծվում են ինֆորմացիայի տեսության գործիքների և մեթոդների միջոցով լուծված գաղտնիության տարաբնույթ խնդիրներին նվիրված մի շարք հրապարակումներ: Հետազոտությունները ցույց են տվել, որ ինֆորմացիոն տեսական մեթոդներն արդյունավետ են խնդիրների լայն շրջանակի համար՝ անանունությունից մինչև դիֆերենցիալ գաղտնիություն:

Բանալի բառեր՝ Մեծ տվյալներ, անանունացում, դիֆերենցիալ գաղտնիություն, Էնտրոպիա, փոխադարձ ինֆորմացիա, շեղում:

Роль теории информации в области конфиденциальности больших данных

Мариам А. Арутунян[1] и Карен А, Мастоян[2]

[1]Институт проблем информатики и автоматизации НАН РА
[2]Гаварский государственный университет
e-mail: armar@sci.am, kmastoyan@yandex.com

Аннотация

Защита конфиденциальности при работе с большими данными - быстрорастущая область исследований. Первым подходом к конфиденциальности был метод анонимности. Недавние исследования показали, что просто анонимные наборы данных могут быть легко атакованы с точки зрения конфиденциальности. Позже была предложена дифференциальная конфиденциальность, которая оказалась наиболее многообещающей. Компромисс между конфиденциальностью и полезностью опубликованных данных, а также другие проблемы, такие как наличие метрик для сравнения различных способов достижения анонимности, относятся к сфере теории информации. Несмотря на наличие в литературе ряда обзорных статей, возможностям методов теории информации не уделялось должного внимания. В этой статье мы даем обзор новейших методов теории информации для обеспечения конфиденциальности. Анализируется серия публикаций, посвященных различным проблемам конфиденциальности, решаемым с помощью инструментов и методов теории информации. Исследования показали, что теоретико-информационные методы эффективны для широкого круга задач, от анонимности до дифференциальной конфиденциальности.

Ключевые слова: Большие данные, анонимизация, дифференциальная конфиденциальность, энтропия, взаимная информация, искажение