WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n3.15
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Is Data Anonymization an Effective Way to Protect Privacy or Not

Yiping Han, Xinqian Lu

Nanjing Military Representative Bureau, Armament Department of the PLAA, Nanjing, Jiangsu, China

ABSTRACT

This paper examines whether data anonymization is an effective method for protecting personal privacy. With the rapid development of the Internet and artificial intelligence, data has become a key driver of modern societal development, but it also raises ethical and technological challenges regarding privacy protection. Data anonymization protects sensitive data by encrypting it and removing personally identifiable information, aiming to reduce the likelihood of identifying individuals within a dataset. The article analyzes the benefits of data anonymization, including the protection of personal privacy, facilitation of data sharing and transactions, and enhancement of data value utilization, while also highlighting the risks associated with data anonymization, particularly the potential for de-anonymization techniques to re-identify personal data, thereby threatening privacy. The study emphasizes that, despite the risks of data misuse, the rational use of data can bring significant positive value to society. The paper concludes that data anonymization itself is not the problem; the real threat lies in data de-anonymization. To maximize benefits, data anonymization should be used rationally, and risks associated with data de-anonymization should be mitigated through various methods. The article suggests that data collectors should prioritize the protection of sensitive data, and regulatory bodies should strengthen the protection of personal data privacy, adopting technologies such as differential privacy to reduce the risk of data correlation attacks.

KEYWORDS

Data Anonymization; Privacy Protection; De-Anonymization; Data Sharing; Data Privacy; Differential Privacy

1. INTRODUCTION

Today, almost everyone handles and deals with a large number of potential or obvious data as a result of the rapid development of the Internet and artificial intelligence. Data has become one of the most significant driving forces behind the development of modern society as we all know. Researchers gather data from different fields to conduct research. Google's database contains your complete history of searches. Facebook stores a huge amount of information about its users, including their actions, comments, photos, and likes. This data determines which news, movies and ads we see. Additionally, it displays which of our friends' posts show up in our social media feed streams and which of our potential partners appear on our dating apps. However, most of the data involved here is relevant to our personal privacy. How to protect our personal privacy is becoming a serious ethical and technological problem. To solve this problem, personal data anonymization has been proposed. Especially driven by big data, the demand for anonymization in data mining, data transaction and data opening is higher and higher. However, is anonymizing personal data the perfect method to protect our privacy and is it a valuable behavior from a consequentialist perspective?

Data anonymization is a way of protecting private or sensitive data by encrypting data and removing data that exposes personal information. The aim of anonymizing personal data is to reduce the possibility of identifying individuals in a data set. Companies generate, store and process large amounts of sensitive data in their normal business operations. Technological advances flourish because relevant information is found in data generated and shared across sectors and countries. However, in order for shared data to be useful without compromising the identity of the compiled client in the database, anonymization must be used.

2. THE BENEFITS OF PERSONAL DATA ANONYMIZATION

"According to Recital 26 of GDPR: The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."

Once personal data have been anonymized, they are no longer considered personal information. Data processing of such a nature is therefore not covered by the Data Processing Act. It seems to be a great opportunity for those organizations or corporations which have sufficiently robust and securely anonymized data sets to do researches, make transactions and so on which can maximize the overall happiness with these anonymous data. Data that is anonymous is not subject to additional protection measures to ensure its security under the regulations above.

Firstly, during the process of data sharing, data anonymization plays an imperative role in protecting personal privacy from the source. With the wide application of the Internet, the age of data and information provides us with a more open society. In everyone's daily life, more and more scenarios need to be driven by shared data. Around the world, data sharing has become an irreversible trend, and there will only be more forms and channels of data sharing in the future. So, when sharing data is the trend, how to protect users' personal information and privacy from being violated? Data anonymization is the best protection measure in information exchange. That is also the reason why GDPR is so restrictive on the process of anonymized data. In terms of the benefits of data collection and sharing, data anonymization provides security protection measures for users worldwide to maximize the interests of all. For example, the Coronavirus has been carrying on for nearly two years. With COVID-19 outbreaks around the world, governments have chosen to share their national data, such as the number of infections and deaths, with WHO. This process of data exchange is critical and necessary to understand how the virus spreads, where it comes from, and how to deal with it. Nevertheless, sharing each country's epidemic data is also extremely risky, since the data may contain sensitive patient information. For research purposes, personally identifiable information is not useful. Anonymized data can protect patients' personal information and facilitate statistics and research.

On the other hand, anonymized data can create much more value for private and public institutions than raw data. According to GDPR, anonymous data is given greater rights to be used. First, when data collectors process anonymous data, they no longer need to obtain user consent. When a company wants to process a user's raw data, the user may say yes or no. It takes companies more time and cost to process through the procedure to get consent and companies may still get uncompleted datasets. Anonymized data solves this problem perfectly. That means companies can analyze user needs or improve their personalization services more efficiently, for instance. In the financial sector, companies are always looking for better ways to reduce risk which can be realized by analyzing data. However, according to GDPR, the analysis of the data requires the customer's consent. The sufficiently anonymous data doesn't. Financial firms can derive more insight and value from their data. For example, when large-scale anonymization of customer data occurs, financial firms can develop and train systems that can detect suspicious behavior and transactions to improve fraud detection and prevention. The next benefit of anonymous data is that the data collector can use the anonymous data for other purposes. This doesn't have to be the original purpose when the data was

gathered. If anonymized data can be shared, exchanged, or even sold, it will not only benefit the data collector, but also intuitions who require a great amount of data for any reason, but are unable to obtain it themselves. Hundreds and thousands of patient information are collected by the hospital each day including their names, addresses, and medical situations, which are commonly required by medical labs and companies. Laboratories and medicine companies need to develop new drugs or treatments based on a large number of patients' data. However, they don't have the capacity to collect so much data. Hospitals become reliable providers of data. Generally, anonymous data can be legally used by a medicine company without requiring the anonymous portion to be disclosed. In the future, it could improve medical treatment, which is a win-win situation for both patients and companies. Also, the anonymized data can be exported internationally. This means that anonymized data need not be restricted to a particular country or region. As mentioned above, countries share anonymous epidemic data with WHO, and many for-profit organizations can benefit from this regulation. For example, many multinational retail companies, such as P&G, Wal-Mart and so on, need a large amount of consumer anonymous data in different countries to analyze different consumption habits to make marketing strategic plans and better business decisions in different countries. It also can avoid losing market share which may be caused by different climates, races, religions and so on.

In terms of consequentialism, data anonymization maximizes the overall convenience benefits. By anonymizing data, the exchange of data is more secure and widespread. This maximizes not only the benefits of data collection, but also the value of the data on top of it.

3. POTENTIAL RISK: DE-ANONYMIZATION

Based on the benefits listed above, it appears that data anonymization is important for individuals, private and public institutions. Consequentialism holds that ethically right actions are those that benefit people. Yet when people want to maximize their profits, data anonymization is no longer a technology to protect people. De-anonymization comes into play here: this is a brand-new method. This technology brings more potential risks to people.

De-anonymization refers to a data mining strategy in which anonymous data is cross-referenced with other data sources to re-identify anonymous data sources. In short, it is possible to re-identify already anonymous data sources by sufficiently comparing several data sets. In 2015, Dr. Latanya Sweeney, a researcher in the field of data anonymization and privacy, conducted a study to "de-anonymize" medical records protected by HIPAA (Health Insurance Portability and Accountability Act) in Washington, DC. In that state and many others, companies and individuals can buy anonymized medical record data. Sweeney buys the data through legal channels, including records of nearly all hospitalizations in the state during a year. These records provide extensive details on diagnoses, surgeries, attending doctors, billings and more. The records are anonymous because they do not contain the patient's name or address, but includes ZIP codes, which are postal codes in the United States. Sweeney then searched the archive for stories containing the word "hospitalization" in Washington State newspapers since 2011 and found 81 qualifying records. By comparing the content of the articles with an anonymous database, there were 35 of the articles had a unique medical record that accurately matched the articles in the database. The news reports clearly included the names of the patients, successfully de-anonymizing the 35 patients.

When anonymized data is de-anonymized, personal information can be collected unprotected. As a result, people in daily life receive nuisance phone calls, sales information, and more seriously, stolen credit cards and other financial information. Even their personal safety may be threatened. From the perspective of consequentialism, the results of information de-anonymization do not maximize overall happiness. At the expense of the majority of users, a small number of people or organizations benefit from the process of de-anonymization. This is definitely not a morally right thing to do. De-anonymization is a negative derivative of anonymized data, which is used by a subset of special-purpose groups to maximize their interests. Much of the de-anonymized data is used to violate privacy

illegally or unethically, but it is not the negative consequences of anonymized data but the misuse of de-anonymized data.

So, what makes de-anonymization so easy? In some cases, there are legal requirements for data anonymization, such as HIPAA's requirement for personal medical data. However, the protection that HIPAA offers is not as strong as most people think. Similarly, the GDPR imposes strict restrictions on the use of data that can identify personally identifiable information, compared with fewer restrictions on the use of anonymous data. Various data security regulations of different governments and industries seem to focus primarily on data that can identify personally identifiable information, while ignoring the risk that anonymized data may also reveal privacy, especially at the point where technical measures are inadequate to ensure that anonymized data cannot be de-anonymized.

In addition, although some companies have adopted "data anonymization" as part of their strategy, it has not been well implemented. Unlike Google and Facebook, for example, Apple deliberately collects less data because having large amounts of data could make the company an easier target for hackers. At the same time, Apple strives to anonymize the data it collects and does not resell user data. These steps are courageous and should be encouraged. Unfortunately, studies have shown that much anonymized data can be easily de-anonymized, especially when multiple data sources overlap to some degree, and any information that distinguishes one data source from another can be used for de-anonymization. And as more data becomes public or leaked, it will become easier to de-anonymize. Most types of data cannot be completely anonymous permanently using traditional anonymization technology, which focuses on protecting personal identity information. Location data is highly unique and therefore more difficult to anonymize.

4. SUMMARY

However, there is the fact that cannot be denied, despite the risk of data abuse, the rational use of data can bring significant positive value to society. We want medical researchers to create new medicines and treatments, our houses to automatically be adjusted to a comfortable temperature, and Google maps to tell us the traffic situation. Benefits of big data are preferred but without the risk of de-anonymization. The truth is that people have to make trade-offs. Consciously or unconsciously, people have had to give up some privacy in order to make their lives more convenient, and may have to give up more in the future.

Data anonymization is not only a means of avoiding the burden of regulatory requirements outside the scope of the data protection law. It was originally intended to reduce the privacy risk associated with the disclosure of personal data. Enterprises that take anonymization measures can provide users with more security guarantees. They can let users know that the information they collect does not use identifiable data when used for big data analysis. This will enhance users' trust and sense of security in big data applications. The use of anonymity should be governed by legality and compliance in order to ensure anonymity serves as more of a security barrier than a protection from data abuse. From the utilitarian point of view, data anonymization brings people more secure information protection and gives institutions a reasonable and legal scope to utilize anonymous information. In general, information anonymization is a process that brings benefits to people, while the real threat to people's privacy is data de-anonymization. The disclosure of privacy information and the abuse of data caused by data de-anonymization have such a terrible impact on people. Based on this consequence, data de-anonymization must be morally wrong, but not data anonymization itself. Utilizing data anonymization rationally while reducing data de-anonymization through various methods is the most effective way to maximize benefits.

Protecting sensitive data and preventing unauthorized access must be a top priority for every data collector. At the same time, regulators should continue to strengthen the protection of personal data privacy. For instance, GDPR encourages companies to store less data and do their best to anonymize

stored data, even if it's not perfectly effective. Similarly, each party involved in data collection and storage should be kept up to date with the latest privacy technologies. Strategies such as Differential Privacy (a cryptographic-based privacy protection technique), where random noise is added to a database before publication, can help reduce attacks based on data correlation. While this method does not reduce the usefulness of the data, it interferes with the calculation process with noise, causing the original data to drown. In other words, people who de-anonymize data by cross-referencing different databases cannot be reconstructed from the original data. Both Apple and Google have made a lot of efforts in studying differential privacy policies.

All in all, people need to be honest about the value and risks of data and treat them carefully. On the one hand, big data is much more powerful than we thought to bring positive improvements to our daily lives. On the other hand, the existence of large amounts of data is a privacy risk. Only when data is protected and used safely and effectively, people are the real beneficiaries.

REFERENCES

[1] The anonymization of personal data. (2021). Retrieved 20 November 2021, from https://www.datatilsynet.no/en/regulations-and-tools/reports-on-specific-subjects/anonymisation/?print=true

[2] FRANKENFIELD, J. (2020). De-Anonymization Definition. Retrieved 20 November 2021, from https://www.investopedia.com/terms/d/deanonymization.asp

[3] FRANKENFIELD, J. (2021). Data Anonymization Definition. Retrieved 20 November 2021, from https://www.investopedia.com/terms/d/data-anonymization.asp

[4] Lubowicka, K. (2018). The Ultimate Guide to Data Anonymization in Analytics. Retrieved 21 November 2021, from https://piwik.pro/blog/the-ultimate-guide-to-data-anonymization-in-analytics/

[5] Data Anonymization: 7 Essential Use Cases. (2019). Retrieved 21 November 2021, from https://www.cloverdx.com/blog/data-anonymization-7-essential-use-cases

[6] DeCew, J., 2018. Privacy (Stanford Encyclopedia of Philosophy/Spring 2018 Edition). [online] Plato.stanford.edu. Available at: <https://plato.stanford.edu/archives/spr2018/entries/privacy/> [Accessed 5 December 2021].

[7] Sweeney L. Only You, Your Doctor, and Many Others May Know. Technology Science. 2015092903. September 28, 2015. https://techscience.org/a/2015092903/