The New Social Necessity – Data Privacy

Aaloka Anant* and Ramjee Prasad

CTIF Global Capsule, Aarhus University, Herning, Denmark
E-mail: aaloka@anantprayas.org; ramjee@btech.au.dk
**Corresponding Author*

Received 18 September 2020; Accepted 30 November 2020;
Publication 26 January 2021

Abstract

Data Privacy is a more prominent necessity than ever in the world we live. The world is so much connected with use of technology than ever in the course of history. Intelligent Machines and technology with thinking power of its own is defining new standards of business operations. Increasing dependency of new generation of people on social technology platforms and digital economy, comes with a big demand from economic and social structure to not only adopt but embed technology. In the same way technology to adopt the social practices like a central governance, democratic values and privacy interlaced with this structure. This paper presents the argument that the protection of privacy of the data subjects is a must have in order to retain the current social structure. We must have a way of using information out of the vast datasets without impacting individuals, indicating a major shift in technology development processes in practice. The new social necessity drives the new technology necessity.

Keywords: Data security, data privacy, anonymization, privacy protection, social benefit, software development practices.

1 Introduction

Data privacy is the central concept for protecting the data of an individual. Any data generated by an individual may be used to identify the individual

Journal of Mobile Multimedia, Vol. 17_1–3, 273–298.
doi: 10.13052/jmm1550-4646.171314

and manipulate his/ her behaviour. No individual wants to reveal any personal information except if it's for their own benefit. Every individual share his/ her personal information with government departments, companies where they buy things from, different events where they register, different social platforms where they interact with others, different service providers where they get services and the list goes on. Without providing the needed personal information, an individual cannot expect to get needed services.

This data contains information which may be misused to manipulate individual's behaviour. To prevent such misuse, data security is a must have. Data protection and privacy protection sometimes sound synonymous as they are based on similar approaches. Though data protection is much more wider topic and privacy protection is very specific topic, which may or may not depend upon data protection techniques. Data protection is more about protecting the data from un-authorized access and ensuring integrity. Though Privacy Preservation is more of enabling access to data, while preserving the privacy of the individual, whose data is concerned.

Data security has a different meaning in different context and also an evolving meaning with time. In the early days before the development of information systems, limited resources were available and most of the data would reside on paper. Data security would mean protecting physical access to these papers and storage mechanisms. With the advancement of internet and the ability to transmit data across the globe with ease, evolved many methods including encryption (different methods), hashing (different methods) and many more techniques.

Despite the advancement in technology, there always has been the need to do more. Data has always been vulnerable. Data has always been key to make intelligent decisions and the lack of it. Not only businesses and Governments have been the consumer and creator of data but every individual. Data has been a vital asset even for the individuals, with the advancement of technology and communication devices like mobile phones. With the increasing generation of data and increasing digitalization, data utility cannot be ruled out only due to concerns on security and privacy. Though at the same time, it's important to use data wisely, in order to avoid scandals like US elections and BREXIT and the involvement of firms like Cambridge Analytica.

This paper carefully analyses the advancements in privacy protection technology and at the same time highlights the need of a Privacy Protection framework. There is no development or technology framework which guides

software design and development to protect privacy of data subjects. Data privacy as a practice is still evolving with technology and is being adopted on a voluntary basis with no technology guideline available at large to guide software development.

2 Why Privacy Preservation Matters

Every individual love to have a personalized service. Its delightful to have a special treatment. Individual is presented with a consent or contract clause, which he/she rarely reads through before providing his personal information on a website or in a physical store, where it is converted into digital format. In most cases, it's the urgency of receiving a service, or the word of mouth, which prompts an individual to share his/ her personal information. Benefits may not be limited to any monetary benefit or a quantifiable incentive, but can be something emotional, which an Individual receives in return of sharing his/ her data to "other party".

The topic of privacy preservation is getting more and more important because now we have advanced technologies to handle large data-sets. These technologies increase the potential of what can be done with these data sets. The benefits can be skyrocketing as well as the dangers.

2.1 Advancement of Technology

Hadoop usage to handle large data sets and derive information increased in the last 10 years. Machine learning concept, which were in research and needed huge investments, are now available for anyone to experiment with small investments, rather than the NASA and niche organizations with big budgets in the past. Processing speed of hardware and software has reached inflexion point, indicating the change of era of data processing technologies.

2.2 Growing Power with Profit Making Organizations

Organizations have outgrown governments in size and budgets. For example, Microsoft, Amazon, Facebook, Apple and Google earn more than the entire GDP of several individual countries. These organizations work for profit and have deep pockets to invest in technologies for artificial intelligence rather than Government organizations. There is no technology governance across the globe and technology companies compete for making getting a larger business in digital economy.

2.3 Awareness in Masses

Once individual has given his personal information and it is converted into data sets, which can be subject to analysis by technology, "other parties" gain the power to manipulate individual's behaviour. In a democratic environment, where every individual is free to give his information or not, can we really stop by making regulations, which enable individuals, to have the right to share their information?

Practical answer to the above question is a prominent "no". The individual is not as empowered as the group of individuals like companies, government, trusts and other parties. These "other parties" have much more resources and power than individuals. Individuals are not even trained or aware about the risks of sharing their personal information on different digital platforms, or even physical locations (stores etc.), where its converted to digital form.

Moreover, individuals in several countries do not even have education on data privacy and privacy protection. They are more vulnerable as the technology has reached in every nuke and corner of the earth via mobile phones, but the awareness on misuse of information, not so much. For a poor subject, who use a mobile device for a service like making payments, and their data is misused, we are looking into a global catastrophe. Also, there is a high risk of manipulating human behaviour with targeted attacks in a geographic region with misuse of personal information of individuals. Manipulating elections and mobilizing mobs for protests can be some of the already seen examples.

2.4 Statistics from Data Breach Monitors

As per one of the most popular web browser Firefox monitors [1], 2019 saw a lot of data breaches and personal information leaked. 2 billion passwords exposed in one single year across the globe. Is there any service with any authority, which informs the individuals, whose data has been breached? The simple answer is, "no mandatory service of that nature exists". Even when such breaches are identified, they are published in general media and the individuals whose data has been breached are not notified. With the most under – privileged internet user in mind, most likely his/ her data has already been breached and he is fully exposed for manipulation by other parties. Statistics of data breach from various agencies are presented in Table 1 below.

As per the latest report from Norton presented in Table 1, a leading company in cyber security space, mega breaches grab headlines, but hundreds of less familiar data hacks also increase the risk of identity theft [2].

Table 1 Data breach statistics published by several agencies

Fact	Agency
First 6 months of 2019 saw 4.1 billion records exposed	Forbes
Cyber attaks considered in top 5 risks to global economy	World Economic forum
Facebook had 540 million records exposed on Amazon cloud server	CBS
There was an 80% increase in number of people affected by Health data breaches in 2019 compared to 2017	Statista

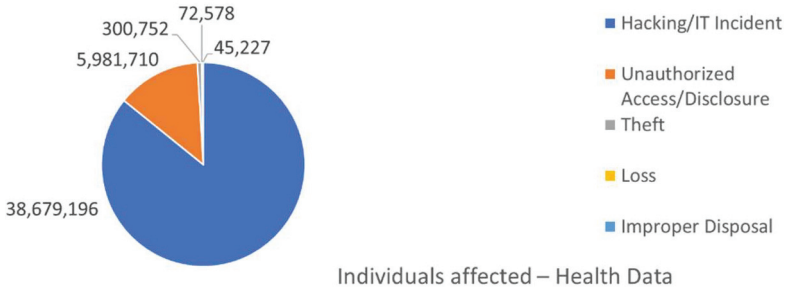


Figure 1 Data breach statistics collected from HIPAA website.

These breaches have been in the area of Financial data, Entertainment data, Healthcare data, Education data, Government data and other business data.

As per HIPAA protection authorities, which tracks health data security in USA, over 40Million individuals can see an adverse impact, due to misuse of their health data. Below chart Figure 1. represents statistics derived from data on HIPAA website.

3 Increasing Awareness in Society

People in general are becoming more and more aware of their privacy rights. Several cases of violation of integrity of individuals has been reported. There are no safeguards in place on a global scale but only in some countries to protect the individual like GDPR in European Union which has 27 countries. Even in EU, one could run targeted marketing campaigns, make phone calls to people, or send them targeted mails, emails, letters by easily finding loopholes in regulation before 2018 June. There have been companies, whose business was to collect and manage individual's information and sell them to other companies, who can run targeted marketing campaigns against those individuals.

Breach of privacy for an individual can cause financial loss as well as other damage in reputation or other moral damages. Cyber bullying with such private data of an individual may lead to manipulation of an individual's behaviour and may even lead to suicides with loss of life itself. There have been several legal cases, where individual was arrested based on tweet alone. In one such case, the tweet was found suspicious by police and the individual was arrested as a precaution for crime prevention, by using his address and location information from the device he used to tweet with and his face for identification and no other evidence [3].

3.1 Large Global Impact

Negative aspect of compromising privacy: The authorities across different countries come to action, when there are mass breaches, which are exposed. And more when the government sees a compromise in security, for example the US authorities cracking a whip on the company Huawei. The global impact of the data privacy breaches can be huge.

Positive side of allowing privacy preservation: Organizations like wikileaks which published the information from several whistle blowers in public platform, exposed several wrong doings. These would have never been exposed otherwise to society. Such platforms have a huge global impact, and the need is already there for such measures on a global scale to facilitate protection of individual's privacy at the same time, usage of data in such a way that the vulnerable individuals are not impacted. Whistle blowers came to wikileaks and other such platforms as it provides protection to individuals submitting such information. Journalism is a major area, where individuals are impacted at large, and it's the responsibility of the journalists, to safeguard the informer who bring value to society by sharing the story.

3.2 Monitoring by STATE

In China, the drive to safeguard individual's data, has been taken up in a different way by the Chinese government. Instead of giving the freedom to the individual, Chinese government is taking all the data of individual to do surveillance. In fact, a social credit system is being designed to benefit good citizens and discourage bad citizens based on criteria set by the government. Government is using all the data including voice calls, facial recognition via surveillance cameras, social interactions by individual and all possible digital presence. On one hand it claims to crack down on the ill effects of data

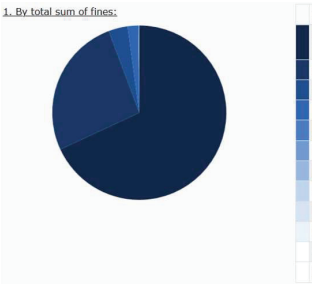
breach and misuse of personal information, for example, this system would enable the government to more effectively crackdown rumour mongers, data thieves, unauthorized VPN connections and ill usage of data. On the other hand, it would give immense power to Government of China to manipulate the behaviour of individuals based on the objectives of the Government. And more so the Government officials who are in charge of running and maintaining that system of monitoring and surveillance [4].

3.3 Europe Leading the Change

Other extreme to China is Europe, which is giving full data privacy to individuals via its General Data Protection Regulation, GDPR [5], which became law across European Union since June 2018, with fines for breach, ranging up to 4% of the gross revenue of the firm responsible for the breach of the regulation exposing data of individuals. In fact, GDPR, even forbids any non-necessary data collection by any company, which is not mandatory to provide a given service to an individual. Individuals have a forum to complain against such request by any company, which would then be looked into and the company may be fined, or at least directed to correct its data collection practices to provide a particular service to individuals. California Consumer Privacy Act, CCPA is the law being formulated in California US and coming into effect since 1st January 2020 to provide protection to individuals data [6].

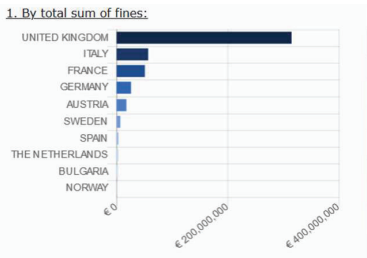
A brief analysis of the GDPR fines clearly suggests that the majority of companies in Europe are not really aware of the right measures to ensure information security. Any company having any business in EU, need to comply with these regulations. This includes any website, which is available for people to see in EU countries. If there is any data collected by those websites of the user viewing the website, they have to comply with EU regulations. Around 80% of the fines, by value, amounting to over Eur 330 Million were imposed on 83 entities for "Insufficient technical and organizational measure to ensure information security". This is presented in Figure 2 below. Hence around 25% of fines, by count, cumulated to 80% of fines by value [7].

Also the distribution of fines across the countries in Europe is interesting as presented in Figure 3. This reflects clearly that the fines are more for companies which have higher revenues, the underlying principle in GDPR to relate the fines to the earning of the company. There are 37 fines in Romania, but the total value is not enough to compare to fines with countries in the top ten list in terms of value of the fines. This also reflects the value such



Violation	Sum of Fines
Insufficient technical and organisational measures to ensure information security	**€ 335,201,807** (at 83 fines)
Insufficient legal basis for data processing	**€ 128,859,040** (at 155 fines)
Non-compliance with general data processing principles	**€ 17,571,565** (at 64 fines)
Insufficient fulfilment of data subjects rights	**€ 9,534,197** (at 41 fines)
Insufficient fulfilment of information obligations	**€ 568,305** (at 20 fines)
Insufficient fulfilment of data breach notification obligations	**€ 220,725** (at 9 fines)
Lack of appointment of data protection officer	**€ 136,000** (at 4 fines)
Insufficient cooperation with supervisory authority	**€ 135,211** (at 15 fines)
Insufficient data processing agreement	**€ 14,380** (at 2 fines)
Insufficient cooperation with supervisory authority	**€ 4,400** (at 1 fines)
Unknown	**€ 500** (at 1 fines)
Insufficient fulfilment of data breach obligations	**€ 286** (at 1 fines)

Figure 2 GDPR fines analysis – by fine type (Source: enforcementtracker.com).



Country	Sum of Fines
UNITED KINGDOM	**€ 315,310,200** (at 3 fines)
ITALY	**€ 57,231,000** (at 28 fines)
FRANCE	**€ 51,350,000** (at 6 fines)
GERMANY	**€ 26,377,925** (at 26 fines)
AUSTRIA	**€ 18,070,200** (at 8 fines)
SWEDEN	**€ 7,085,430** (at 6 fines)
SPAIN	**€ 3,683,910** (at 132 fines)
THE NETHERLANDS	**€ 3,490,000** (at 6 fines)
BULGARIA	**€ 3,210,690** (at 20 fines)
NORWAY	**€ 1,032,060** (at 10 fines)

Figure 3 GDPR fines analysis – by country (Source: enforcementtracker.com).

a regulation is trying to bring into the lives of people. At the same time, promoting innovations and not taxing small companies which deal with data, by imposing big fines to shut them down, but for sure, giving enough blow to warn them and take care of individual's data security and privacy.

4 Why is it so Difficult to Preserve Privacy

Privacy is not an attribute of data. There can be different data elements which can be considered as private. Personal and sensitive are two terms, which can define what is private. For example, name of a person is private information for the individual, though it may or may not uniquely identify an individual. In general, any attribute, which can be used to uniquely identify an individual can be considered a private information. Another term relevant in this context is sensitive information. When some information can be sensitive to impact an individual, even though it may not be something personal to an individual, it can be considered private and something which must be preserved to avoid any misuse of such information to harm the individual.

This broad meaning of personal and sensitive information makes the challenge of privacy preservation difficult and even impossible to realize with a pure technical solution.

4.1 Volume of Data

According to an IDC whitepaper by David Reinsel "Mankind is on a quest to Digitize the world" [8, p. 3]. Enterprises are increasingly storing this data. Individuals benefit from the services provided by these enterprises, for example Google provides 15 GB of storage to users for free for storing photos. Enterprise in turn benefit from this data under contractual agreements with the individuals. Overall data generation is on the endpoints with the mobile devices and PCs used by individuals and the new IOT devices as machines, and accounts for the huge growth in data to over 175 Zettabytes by 2025. A Zettabyte for clarity is 1000 Exabytes, which is 1000 Petabytes, which is 1000 Terabytes, which is 1000 Gigabytes. IDC is talking about 33 ZB already in 2018, which means something like 33 trillion 1 TB computers, something like 5000 1TB computer worth of data for every living being in 2018. Even though individuals may not have a single computer in their hand, but the data generated in relation to different interactions of these individuals within society, and stored with repeated copies, make this huge data set.

4.2 Types of Usage of Data

In terms of Enterprise data, companies, categorize less frequently changing data, which has information on people, products etc. as "master data" and more frequently changing data or data with a timestamp as transaction data. Transaction data may not reveal a lot of personal information except if combined with master data. Though it may contain attributes like name, card details etc, which may accurately reveal a person's identity. Data engineering or feature engineering may be used /misused, to derive personal or sensitive information from a large data set with high accuracy, even though any personal information may not be present in the data set.

4.3 Storage Location of Data

Where is this data stored, is another very significant fact to establish the need of data privacy. As per the same IDC [8, p. 10] report, the data is increasingly being stored in the public cloud setup. This is a major factor as earlier trend was the storage of data increasingly at the edge, the device where

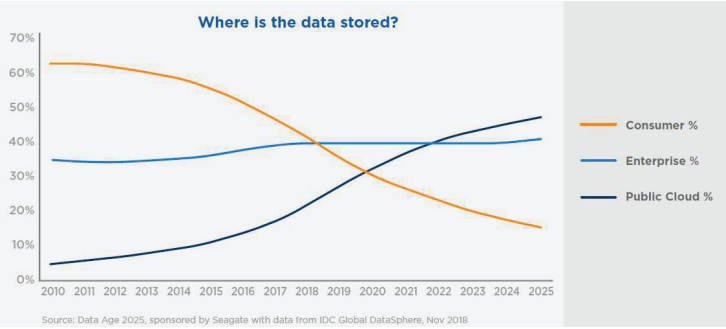


Figure 4 IDC: Where the data is stored (trend and forecast) 2010–2025.

it is generated, for example a mobile phone. With increasing size of data and improved reliability of public cloud environment and increased upload and download speeds with 4G and 5G networks, data storage in public cloud is becoming a commonplace. Figure 4 below from the IDC DATA Age 2025 represents the shifting trend of data storage.

With the storage in public cloud, limited in features for individuals, and access available to the cloud storage offering company, creates a unique challenge to data privacy. For example, a company offering free storage for cloud photos, can use the photos of an individual to improve its face recognizing algorithms. The individual is not even aware of the same and is bound by a contract for it, but at the same time, he gives up his data to the company offering free storage. The company can make good use of data in improving its algorithms, which they can later use for other purposes and make money. This is good or bad, ethical or not ethical is a related question, but not currently addressed under any framework.

4.4 Technology Used for Storage of Data

In terms of structure of the storage of the dataset, different technologies, enable different types of storage. The Enterprise dataset is primarily in structured form as relational tables. Such a storage facilitates easy storage, easy retrieval and high concurrency in order to enable real-time operations. ACID (atomicity, consistency, isolation and durability) compliance is one of the standards, which certifies technology being used to manage Enterprise Data. Other technology also termed with BASE (Basically Available, Soft state, Eventual consistency) [9] compliance is perfectly valid and used for many other uses and storage of data like social media. Due to variation in

basic methods on storing data, these technologies have different methods of querying data. The data types, in use are also different. Images and videos are not so commonly stored in Enterprise data, though they are the most common elements to store in data in social media. This is also a reason why social data is growing very high in volume and footprint across the globe.

4.5 Structure of Data

The data structure is a good topic when it comes to anonymization and different methods to find the most effective method.

- Structured data is primarily data with a defined structure, primarily in tabular form. This dataset is stored in databases mostly relational and has well defined authorization and access management. This is in fact the most widely used type in Enterprise world.
- Unstructured data is data stored in a running format. For example, free text. This can be used to represent many forms of data like media, images and any other form, which is not structured.
- Semi-structured – this term is also used in some places, where a structure can be easily derived/ seen in the stored data, though data is not stored in a tabular format. For example, an email. Some argue that email is an unstructured data, but some argue it is structured as it has a header, body and defined format. This category does not have a clear demarcation and is not widely used.
- Graph data – this data type is unique and evolving in how it is stored and used in database. Even though it has a structure, it is meaningful only in a certain context of usage. There are specific technologies, which guide the use of this data type, and provide a lot of convenience to dealing with locations, specially when using this data type. This data type is pretty clear in what it represents and even though it has no personal information on its own, it is very significant in the world of mobile technology.

4.6 Graph Datasets and Location Information

With the GPS enabled on the mobile phones, or otherwise as well, due to nature of the technology of mobile communication, the accurate location of the mobile phone is always available with the telecom network. It can be a dataset, which is very much revealing of personal information about an individual. Its usage to find out personal information about an individual is

not unseen, with so many movies being based on the same. Though in the context of usage for privacy preservation, dealing with this data, presents a very unique challenge, when storing or analysing the same, in order to protect privacy. There has been a clear evidence that even though the data is not associated to any individual, if available data points are available for an unknown person, it is very easy to identify patterns and find out who is the individual [10].

5 Primary Factor

In this section we analyse the primary factors, which affect privacy preservation. The digital world is primarily driven by the software programs, which make different applications run. Every application software or online program or any digital entity is guided for operations by the lines of code, which determine the sequence in which an operation would take place in the digital world.

5.1 Why Store Data Once Its Purpose is Complete

Each instance where data from an individual is collected is programmatically designed to process that data and relate it with other data in the application to perform an operation for the individual or the company, which owns the software. Every instance of data storage in any digital form is preceded by data processing. Every instance of data is stored for a purpose. The purpose can be to run the software application in order to benefit the individual or the company which owns the software. Once the purpose is solved, this data is not relevant. Hence the data may be deleted.

5.2 No Information Stored with Data, if it is Privacy Relevant

It is a big task in the technology research on privacy preservation area, as which data is relevant for privacy preservation. Technical solutions are being presented, though it's a very vulnerable topic due to the factors discussed in this article overall. Why is there no control, whenever a software application collects data from an individual or about an individual or any data, which can affect an individual. At the time when data is collected, why is it not marked appropriately, in order to be identified as permitted for further processing or not. With an additional information stored with the data, it would be easier to identify data which needs preservation.

5.3 New Usage of Old Data

With the increasing use of Machine learning and artificial intelligence technologies, the primary purpose is to predict the behavioural patterns and simulate event outcomes in future. This needs the machines/ software to know what has happened in the past and identify patterns. For this purpose, data which was stored in the past and has been deemed to be of no use further, gains a different meaning. This data from past, which may not have any other use, can be used to train a machine, to predict future events.

The problem is not why machine learning or artificial intelligence needs this data, which may be a discussion outside scope of this paper, but importantly, why there is no way to identify which data should be allowed for being processed further and which data should not be allowed for any further processing, in order to protect the privacy of individuals.

5.4 Guidelines Available for Businesses Not for Software Makers

Guidelines, laid out by laws like GDPR, define clearly on anonymization (discussed in next section). Anonymization is needed in order to alter or regenerate data, so that individual can be protected. Such guidelines not only raise a valid concern, they clearly raise a bigger question on how the software applications handle data and why.

There are currently no programming languages, which allow this kind of information to be stored by design. There are no compilers, which check for such standards. In fact, there are no programming standards to ensure that the personal information is dealt appropriately by the software application.

5.5 No Standards or Checks Available for Privacy Relevant Data

It's not the first time, that we have this situation, but this has been a case in the past. For example, information about the bank account, credit cards, financial transactions and other such critical information, which can be misused to cause financial damage, cannot be stored in applications, until they meet certain standards [11]. PCI Security Standards council lays out guidelines on what may be stored and with what standards of security when it comes to information like credit cards or other such information. No such auditing standards exist for privacy relevant data.

Privacy is an attribute which affects sensitivity and emotions of a person and hence can be more damaging than the financial information loss. But sadly, there are no standards defined on how privacy relevant data sets should be captured and stored by different applications.

6 Ensuring Privacy is Preserved

Ensuring that privacy of data subjects is preserved is even a more open topic at the time of writing of this paper. There are no auditing standards defined by law. There are only guidelines available on possible usage of data collected from individuals. Authorities like GDPR do plan to provider certification to authorities which can assess and certify the data controller and data processor. Every company in the coming years in Europe would have to have the certification of its personal data processing. The practice would ensure that data subjects, individuals, which are vulnerable and not able to understand the dangers of losing their privacy would also be protected.

Such measures as in GDPR are not available across any other governance body across the globe. To compare it with existing standards would be something like the accounting standards, ISO standards for quality checks, process monitoring, testing in software industry to name a few. California Consumer Privacy Act is first such law in United States of America. Though it does not yet have detailed regulations on certification and ensuring that the privacy of individuals is preserved except for having fines. A detailed analysis of these regulations is outside the scope of this paper.

7 A Concept to Preserve Privacy-Anonymization

Anonymization is primarily defined as a process for information sanitization the intent of which is privacy protection. Different industries see it differently. For example, medical information about a patient is recorded in public health records[12]. These are not public, but accessible to doctors in the system and many other authorized staff. In some cases, this data has to be published to provide information to public in general. And also, for researchers, who are doing research on a given topic, where a given patient's data may add value. Such specific and general cases, where this patient's information may need to be published are cases, where it is very necessary to understand the privacy concern for the individual. On one hand, the patient may benefit by sharing his data for a research or maybe not. It may so happen that the patient attracts some un-necessary attention, which he/she is not seeking and may be un-comfortable with sharing such information.

7.1 What to Anonymize

Any data set, which can be used to reveal personal information should be anonymized. It may include the data set, where the information is not personal

to an individual but can be used to identify the individual. For example, colour of hair of a person if combined with his address, may be used to identify who the person is. So, the colour of hair of a person, may not be a personally identifiable information, but in a context, it can be used to identify the person. Hence there is no general rule on what should be anonymized. Rather important is to check post anonymization, that the individuals cannot be identified in a process, for which the data has been anonymized, and to ensure that the anonymized data is not used for any other purpose but deleted after its use for an intended purpose, leaving no trace to identify the individual.

7.2 Anonymization Works Even for Sensitive Data

Anonymization intends to alter data in such a way that the personal/ private information is removed from the dataset. Anonymization may be used for any data and not only private data, but sensitive data, which is not so private, but not something which the owner of the data is comfortable in sharing with others. One example of a dataset with no personal information but information, which one would not like to share is a mobile manufacturing company's information on defective mobile phones and returns by its customers in a given year. Only removing privacy relevant data may not be sufficient, but the broad definition of Anonymization, ensures that the individual is not identifiable. Hence anonymization covers not only removal of personal data but any information, which can be privacy preservation relevant.

7.3 Different Methods for Anonymization

Anonymization can be achieved using different technologies as claimed by the developer and researchers on those technologies. Some of those technologies include below methods.

- Differential Privacy – This method is attributed to be primarily developed by cryptographers. In 2006, published work "Calibrating Noise to Sensitivity in Private Data Analysis", may be considered as the foundation of Differential privacy. Differential privacy ensures that once the data has been differentially private, the user of the data would not be able to identify if a given individual's data is in the dataset being analysed or not. The user even if he/she has information about a given individual, whose data is in the dataset, there is no possibility to identify any other individual from the data set. This is one of the most renowned methods for sharing information publicly as per researchers.

The implementation of this concept has been done by several researchers in different geographies, different industries on different data sets.

- k-anonymity – This method is also very much used in relation to datasets, where the data can be grouped into levels of hierarchy. For example, villages being part of a district, districts being part of a state, and states being part of a country. Based on the defined rules, a particular subjects' data can be replaced with a value, which is higher up in hierarchy in order to protect an individual being identified. As long as there are more than enough individuals for a given value of a record identifier, the algorithm can keep doing aggregations. There have been improvements in this algorithm, with l-diversity and t-closeness, being safer from an implementation perspective for this logic to avoid any personal data identification.
- Other methods – pseudonymization, scrambling, masking and cryptography based many other methods are in use today for data anonymization. Basic purpose of all these approaches is to alter the data in such a way, that the individual's data in a dataset, may not reveal any information to trace back the individual, and find out specifically whose data is being presented.

7.4 Differential Privacy Based Synthetic Data Generation

This presents a unique approach in which the data from the original dataset is not taken after anonymization, but a deep learning [13] or another form of machine learning approach or a mathematical approach is taken to identify patterns in the data. Based on the identified patterns in the data, data elements are re-generated to retain those relationships. But the original data is not reproduced in the output. Such a method ensures that there is no way to recreate the original dataset with any accuracy, but at the same time, it is an attempt to keep the data relevant for use by machine learning algorithms, which need the relationship between the different attribute values as an input to infer meaningful output as a part of analysis.

7.5 Can We Really Preserve Privacy with Anonymization?

Anonymization of data leads to alteration of data attribute values in order to protect the privacy of individuals and to avoid tracing back the individual whose data is being observed. Such an operation also results in loss of utility of data. This loss of utility is a primary element in determining which method of anonymization should be selected. Even more evolved methods of data

encryption like cryptography, appear to be appropriate in some of the cases. It is simply the post anonymization usage of data, which primarily drives the choice of method of anonymization, in combination with other elements like the source and type of data.

8 How to Break Anonymization

Every method comes with its limitations and with the advancement of technology, different methods get outdated and need to be re-defined or modified. In case of anonymization as well, there has been several instances, where a completely un-expected anonymized data set has been used to reveal personal information.

8.1 Methods of Attack

Some of the identified methods of attack which are invented and documented work on the basic objective to make it possible to identify an individual from a dataset, where the specific individual is not mentioned. Linkage attack, Inference Attack, Homogeneity attack, background knowledge attack, social engineering attacks are a few of those. Some of them are described below with examples.

- Linkage attack – in this type of attack, the data in an anonymized dataset, is linked with other dataset, where there is more personal information about the individual. If the linkage is successful, the personal information of the individual, which is available in limited form in one place, can be linked to get much more information on the person. For example, date of birth of a person, if combined with the pin code of address of the person, this can form a very unique combination to identify an individual. Hence if data is published by a hospital with only the pin code and address of a person with details on what all diseases are being treated in a hospital for what age groups of people, then if combined with their published voter records, their names can be obtained. And once names are obtained, health information including the types of disease for which the person has been treated can be inferred from the other dataset from hospital, making the anonymized data from hospital, de-anonymized. This can adversely affect the individual.
- Inference attack – [14] In this type of attack the information is inferred from another available information, with a high certainty. This type of attack is done with various data mining methods and data engineering

techniques. For example, if the location of a person can be verified with certainty and it is a location of a home, it can be inferred who is the person. Hence other movements based on this location information, can be used to make an inference about the movement of a given person. There is certainly a need of distorting this information with noise in order to avoid inference attacks.

- Homogeneity attack and background knowledge attack – these are more of data engineering attacks. When there is a lot of homogeneous information, meaning same value for a sensitive attribute in a dataset, with high certainty, this can be assumed to be true for any subject of the dataset. Hence, there is a possibility of de-anonymization, even though the methods like k-anonymity would have been applied. Background knowledge attack is simple as it states. Having a knowledge of some of the data elements/ individuals, if it is possible to derive information about other individuals in the dataset, by using various data engineering techniques.
- Social engineering attack – these attacks can be intrusive attacks, where a subject is prompted to provide personal information based on fake/ simulated tricks. Using this information, other information about the person can be revealed using their social media accounts, or corporate account or other digital accounts. For example, by calling an employee of an organization as the Information technology staff of the same company, attackers, can get access within an organization via his/her user id for that organization. Then the attackers can find out more information about the employee by accessing organization records about the employee. Also, they may gain access to the same information that the employee has in the that organization [15].

8.2 Hacks for Monetary Value

With many of the attacks described above and other attacks, the objective of the attacker is to get information about an individual, which is not published. Using this information, the attacker, may gain a financial advantage, like withdrawing money from individuals accounts, blackmailing individual, to return some favours or bullying in general in the cyber space. There have been instances, where the whole institution has been put on hold by the attackers, who could muddle with data in such a way that the organization or the individual affected had to pay ransom to the attackers in order to get their data back. This was called ransomware attack and affected hundreds and thousands of users and organizations [16] as reported in 2017.

8.3 Research to Benefit Society or Encourage Breaking Rules

Apart from these direct attacks, which are for immediate benefit by the attacker, there are attacks, which were done for not making any profit but only to prove that the method used for anonymization are insufficient. These kinds of attacks are done by intellectuals and researchers, in order to expose vulnerability in the system. At the same time, a lot of data is exposed but it gives an alarm to authorities to deploy other methods of anonymization to save individuals data [17]. This type of attack prevents other attacks and misuse of data and help the practice of data security advance further.

9 Heart of Research Which Makes the Problem Bigger

There is a lot of work and patents on how to interpret information from the images and videos, to make them searchable using voice or text search. Hence this area is of higher interest from technologists and researchers compared to the Enterprise data. As the boundaries between Enterprise data and non-enterprise data vanish, with more and more companies using social media to connect to their customers, the importance of data security and privacy becomes more pronounced.

These new innovations of identifying the name of the person from a photo, or identifying the address where the photo was taken, has a huge potential in how the social media data can be utilized. Just for a hypothetical example, if the individual is uploading every image taken on phone to a cloud storage, the company managing the cloud storage, has a potential to identify the location of the individual and also identify his needs feeding these data sets to machine learning algorithms. This would mean, the company holding this information on images would be in a position to offer a product promotion to the individual which he/she needs, or the company may predict his/ her need as well. This and many other such use cases, make the logical choice between data privacy and the lack of it a very thin line to cross for several organizations.

10 Can We Achieve 100% Anonymization

In a paper submitted in Open Identity Summit in Bonn, Germany in 2019, authors argue that “Anonymization is dead, Long live Privacy”. They advocate a paradigm shift, away from anonymization towards transparency, accountability and intervenability [18]. There have been several papers

published on the topic of anonymization, over 500 papers in 2017 alone. Despite a lot of work in this area, for over 10 years, researchers have found that the top of class methods are still not able to make data fully anonymized, so that the individuals are not traceable back and still retain data utility. There is a constant search for algorithms to brake anonymity and de-anonymize information.

10.1 Use of Data Context to Identify Individuals

Overall, more than technology it is important to understand, what is being dealt with here. Data which is generated by individuals have their own writing styles, their own use of words, frequency of words and variety of words from a vocabulary. To represent the same emotion, different people may use a completely different term. Also, the same words used in a given context and environment/group, may have a very direct meaning compared to usage of such words outside of that environment.

For example, use of the "man with a white hat" can be very specific to a person who always wears a white hat in a given office. Though white hat represents "information" in case of a "six thinking hats" method of discussion. Hence if a text is written as "man with a white hat was responsible for breaking a glass bottle on the way" – may reveal the identity of the person, if the statement is given by the employees of that office in our example. Though if the same statement is made by someone from outside that office, maybe on a road, it may not reveal any person's identity as white hat can be worn by anyone, walking on the road. Based on this example, it is easy to understand that whatever is the method of making the data not identifiable to personal level, it is very difficult to classify data in first place, which data can be personally identifiable, and context to data is very important, which is most of the time, not stored with data itself.

10.2 Data in Applications Used by Companies to Run their Business

Other type of data, which is enterprise data, may not suffer from the same constraints as unstructured data, which is more in the form of text or comments. Enterprise data is more of structured data, which is generated by applications. These applications process personal information. For example an application to generate an invoice at a point of sales terminal in a shop can process card information and attach it to the transaction, thus storing the card information and other details of purchase. This application would

write the data to a database, where it can be stored and later retrieved for reporting purposes by the company, for example to report the total volume of products sold from the store or the total revenue generated by the store or any other information. Later on, in time, if the customer is returning one of the items purchased, the application would fetch the information on the transaction from the database and provide that service of return of product, to the customer. When these data are used by a company to analyse who are the customers of the company, the company can use this information. This data stored with the company can be used by the company for any purpose technically.

10.3 Greed Driving Misuse of Data – No Technology or Lack of it

Data about the customer can be easily misused using technical means for profits. For example, if a person is buying medicines for treatment of a disease, and this data is provided by the selling shop to external parties, then this can be used by other companies to market products to this person or the doctor, manipulating their buying decision. Also, if this data is sold to a political campaign manager, they can prepare customized incentives for this person, this can manipulate their voting patterns in an election. Technology can be used to anonymize this data. Then the owner of the data, which is the business selling medicine, may provide this data to an external company, which can analyse how many patients are in a given area, and may increase stock of medicines. But in no way, this external company would have any way to reach the individual and manipulate his/her buying pattern. Also, if anonymized appropriately, the data can be provided to a political party for example, which can design a campaign to improve the condition of patients in a region in general but would have no way to manipulate the votes of an individual by reaching out to him for bargaining on incentives.

11 Conclusions

Data privacy preservation is a topic of increasing interest. This paper highlights different aspects of privacy preservation. Paper showcases different social factors which have enhanced interest in this topic. Paper highlights the lack of standards in the area of software engineering due to low value given to this topic in the past. Only with increasing focus from legal authorities, and monetary fines being enforced, the Enterprise companies and the software

companies, which create and determine the rules in digital world are taking this topic seriously.

Privacy Preservation of Data needs basic framework and checks to be in place in order to ensure that the problem of privacy preservation is taken care of, right from the first step where data from an individual is collected and stored. The technologies used for anonymization and privacy preservation, used to provide protection to the privacy of individuals, needs to be safeguarded with appropriate framework.

Establishment of GDPR and other such framework form a good basis for creating new data management strategy in existing companies. These frameworks do not have any dependence on technology. Also, they do not recommend or restrict any technology as long as the objectives set under the framework are met. Though the impact of such a framework is yet to be seen as individuals do not need the companies to pay fines. But more, that the safeguards are adopted to protect individuals, without them even worrying about the same.

The analysis of fines in the last 18 months under GDPR clearly indicates that the fines are not for the breach of technology, but for the non-compliance in general. The technology has proven to be enough only if all other safeguards are maintained by the organizations.

References

[1] N. Nguyen, "Introducing Firefox Monitor, Helping People Take Control After a Data Breach," The Mozilla Blog. https://blog.mozilla.org/blog/2018/09/25/introducing-firefox-monitor-helping-people-take-control-after-a-data-breach (accessed Dec. 18, 2019).

[2] "2019 Data Breaches: 4 Billion Records Breached So Far." https://us.norton.com/internetsecurity-emerging-threats-2019-data-breaches.html (accessed Dec. 20, 2019).

[3] J. Wright, "Teen blasted out of Rod Laver after 'Timebomb' tweet," The Sydney Morning Herald, Jul. 08, 2013. https://www.smh.com.au/entertainment/music/teen-blasted-out-of-rod-laver-after-timebomb-tweet-20130708-2pl4d.html (accessed Dec. 18, 2019).

[4] X. Qiang, "The Road to Digital Unfreedom: President Xi's Surveillance State," Journal of Democracy, vol. 30, no. 1, pp. 53–67, Jan. 2019, doi: 10.1353/jod.2019.0004.

[5] P. Voigt and A. von dem Bussche, The EU General Data Protection Regulation (GDPR). Cham: Springer International Publishing, 2017.

[6] E. Goldman, "An Introduction to the California Consumer Privacy Act (CCPA)," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3211013, Jun. 2019. Accessed: Dec. 18, 2019. [Online]. Available: https://papers.ssrn.com/abstract=3211013.

[7] G. Graham and A. Hurst, "GDPR enforcement: How are EU regulators flexing their muscles?," IQ: The RIM Quarterly, vol. 35, no. 3, p. 20, Aug. 2019.

[8] D. Reinsel, J. Gantz, and J. Rydning, "The Digitization of the World from Edge to Core," p. 28, 2018.

[9] W. Vogels, "Eventually Consistent," Commun. ACM, vol. 52, no. 1, pp. 40–44, Jan. 2009, doi: 10.1145/1435417.1435432.

[10] S. Gambs, M.-O. Killijian, and M. Núñez del Prado Cortez, "De-anonymization attack on geolocated data," Journal of Computer and System Sciences, vol. 80, no. 8, pp. 1597–1614, Dec. 2014, doi: 10.1016/j.jcss.2014.04.024.

[11] "pci_fs_data_storage.pdf." Accessed: Aug. 15, 2020. [Online]. Available: https://www.pcisecuritystandards.org/pdfs/pci_fs_data_storage.pdf.

[12] "Sweeney - DAnatoanyym iatySiynstMemedfiocralPDroavtiading.pdf." Accessed: Dec. 20, 2019. [Online]. Available: https://dataprivacylab.org/datafly/paper2.pdf.

[13] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney, "Privacy Preserving Synthetic Data Release Using Deep Learning," in Machine Learning and Knowledge Discovery in Databases, Cham, 2019, pp. 510–526, doi: 10.1007/978-3-030-10925-7_31.

[14] J. Krumm, "Inference Attacks on Location Tracks," in Pervasive Computing, Berlin, Heidelberg, 2007, pp. 127–143, doi: 10.1007/978-3-540-72037-9_8.

[15] Department of Telematics Engineering, ETSI Telecommunication Technical University of Madrid, Madrid, Spain, W. Fan, K. Lwakatare, and R. Rong, "Social Engineering: I-E based Model of Human Weakness for Attack and Defense Investigations," IJCNIS, vol. 9, no. 1, pp. 1–11, Jan. 2017, doi: 10.5815/ijcnis.2017.01.01.

[16] "A brief study of Wannacry Threat: Ransomware Attack 2017 - ProQuest." https://search.proquest.com/openview/e14e6aab226f21a65db72f0b58653572/1?pq-origsite=gscholar&cbl=1606379 (accessed Dec. 18, 2019).

[17] J. S. Yoo, A. Thaler, L. Sweeney, and J. Zang, "Risks to Patient Privacy: A Re- identification of Patients in Maine and Vermont Statewide Hospital Data," . October, p. 62.
[18] J. Zibuschka, S. Kurowski, H. Roßnagel, C. H. Schmuck, and C. Zimmermann, Anonymization Is Dead – Long Live Privacy. Gesellschaft für Informatik, Bonn, 2019.

Biographies

Aaloka Anant is a researcher at CTIF Global Capsule, Aarhus University, Denmark since October 2019. He attained his Post Graduate degree in Enterprise Management from Indian Institute of Management Bangalore and B.Sc in Electronics and Communication Engineering from BIT Sindri in India. Aaloka has held leadership and senior positions in SAP and Honeywell since 2004 and also worked with start-ups like Idea Device Technologies, Movid-DLX, NGeneR and co-founded a non-profit organization Anant Prayas. He is actively pursuing research on the topic of Privacy Preservation. His work focusses on new approaches for privacy preservation of application data and missing technology and structural framework for achieving end-to-end data privacy.

Ramjee Prasad is a Professor of Future Technologies for Business Ecosystem Innovation (FT4B1) in the Department of Business Development and Technology, Aarhus University, Denmark. He is the Founder President of the CTIF Global Capsule (CGC). He is also the Founder Chairman of the Global ICT Standardization Forum for India, established in 2009. GISFI has the purpose of increasing of the collaboration between European, Indian, Japanese, North-American and other worldwide standardization activities in the area of Information and Communication Technology (ICT) and related application areas. He has been honored by the University of Rome "Tor Vergata", Italy as a Distinguished Professor of the Department of Clinical Sciences and Translational Medicine on March 15, 2016. He is Honorary Professor of University of Cape Town, South Africa, and University of KwaZulu-Natal, South Africa. He has received Ridderkorset of Dannebrogordenen (Knight of the Dannebrog) in 2010 from the Danish Queen for the internationalization of top-class telecommunication research and education. He has received several international awards such as: IEEE Communications Society Wireless Communications Technical Committee Recognition Award in 2003 for making contribution in the field of "Personal, Wireless and Mobile Systems and Networks". Telenor's Research Award in 2005 for impressive merits both academic and organizational within the field of wireless and personal communication, 2014 IEEE AESS Outstanding Organizational Leadership Award for: "Organizational Leadership in developing and globalizing the CTIF (Center for TeleInFrastruktur) Research Network", and so on. He has been Project Coordinator of several EC projects namely, MAGNET, MAGNET Beyond, eWALL and so on. He has published more than 30 books, 1000 plus journal and conference publications, more than IS patents, over 100 Ph.D. Graduates and larger number of Masters (over 250). Several of his students are today worldwide telecommunication leaders themselves.