5. Data Privacy Technology for Society

Aaloka Anant, CTIF Global Capsule, Aarhus University, Herning, Denmark aaa.sap@gmail.com
Ramjee Prasad, CTIF Global Capsule, Aarhus University, Herning, Denmark ramjee@btech.au.dk

ABSTRACT

Data Privacy is more prominent than ever in this connected and technology driven world. All information available is derived in one way or the other from data generated by individuals directly or indirectly. This paper presents the argument that we must have a way of using information out of the vast data-sets without impacting individuals. As long as we do not have such a way, there would always be issues, giving rise to misuse of information for someone's benefit and manipulative outcomes. This paper analyzes not only technical but other safeguards, in place which enable an overall benefit of society and individuals with data privacy at its core.

Keywords — *Data security, data privacy, anonymization, privacy protection, social benefit*

INTRODUCTION

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 access of individuals to these papers and storage mechanisms. Coding patterns and secure methods of transferring the data sets were invented. With the advancement of technology and advent of analogue tapes, data storage and reproduction were easier and lead to evolution of methods of protection. Floppy disks came with protection passwords. Furthermore, the evolution continued, with the advancement of digital storage mechanisms on drives to store data in binary format, came many methods including encrypting the whole drives with password. 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 has 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, its important to use data wisely, in order to avoid situations like the case at US elections and the firm Cambridge Analytica.

PERSONAL TOUCH

No individual wants to reveal any personal information except if it is for their own use. Every individual shares 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. Every other organization, which holds individual data, is termed as 'other party' in further text. All this data individuals share is stored and can be used by these 'other parties' for providing goods or services to an individual. To serve the individual in a better way, these other parties can use the data in their operations, where the individual is impacted.

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 is 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”.

WHY WORRY ABOUT PRIVACY PRESERVATION

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 sky rocketing as well as the dangers.

A. INTENT OF USE OF TECHNOLOGY

It is possible to use these technologies and make data useful for business operations, political campaigns or other purposes. If the new technologies are limited to research on nature and doing weather forecasts, or finding out stars and galaxies in universe, individuals may not have a problem, but the moment, the same technologies, which can analyze large amounts of data sets to predict what an individual would do in a given situation, it becomes scary. Scare is not for the reason that technology can assist individuals with the vast amount of knowledge on their fingertips, but scare is by the fact that this capability of technology can be misused to manipulate individual’s behavior.

B. DIGITIZATION

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 behavior. 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 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.

C. AWARENESS IN MASSES

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 the most 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. In addition, there is a high risk of manipulating human behavior 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.

D. WHAT MONITORS SAY

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 who data has been breached are not notified. With the most under – privileged internet user in mind, mostly likely his/ her data has already been breached and he is fully exposed for manipulation by other parties.

Table 5-1 STATISTICS OF DATA BREACH FROM NORTON

Fact	Figures
The number of publicly disclosed breaches.	3,800
The number of records exposed	4,100,000,000
Increase in number of reported breaches in first six months of 2019 vs. first six months of 2018	54%

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

SOCIAL DRIVE CONVERTS INTO A LEGAL DRIVE

There has been severe outrage by technologists over the years and several cases of violation of integrity of individuals, still there are no safeguards in place on a global scale but only in some countries. Prior to 2018, if one would look into technology which can be used to alter voice of an individual, or identify an individual from his/ her data this could be not having any financially punishable offence. Even one could run targeted marketing campaigns, make phone calls to people, even send them targeted mails, emails, letters by easily finding loopholes in regulation on these. There has 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.

Not only financial losses for an individual, but breach of privacy can be in other forms impacting individuals in several ways. 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].

A. LARGE GLOBAL IMPACT

The authorities across different countries came to senses, only 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.

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 to 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 is the responsibility of the journalists, to safeguard the informer who bring value to the whole society by sharing the story.

B. AN EXTREME EXAMPLE OF MONITORING

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 rumormongers, 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 behavior of individuals based on the objectives of the Government. In addition, more so the Government officials who are in charge of running and maintaining that system of monitoring and surveillance [4].

C. 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 Europe 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 Enterprise are not really aware of the right measures to ensure information security. Security of data for individuals have never been so important than after the implementation of GDPR in EU effective June 2018. 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 to EU regulations. Around 80% of the fines by value, amounting to over Eur 330 Million were imposed on 38 entities for “Insufficient technical and organizational measure to ensure information security”. Hence only 27% of fines by count, cumulated to 80% of fines by value [7].

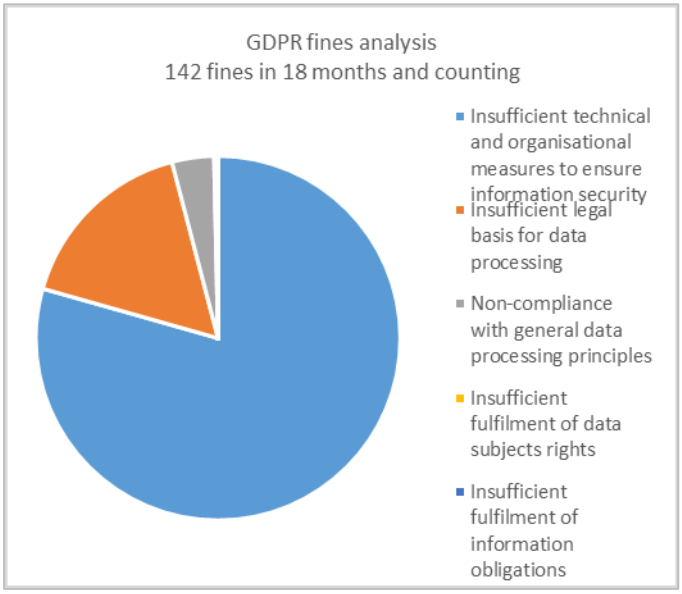

Figure 5-1 GDPR fines analysis (by fine type)

Also the distribution of fines across the countries in Europe is interesting as presented in table 2. 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 11 fines in Hungary, 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 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 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.

Table 5-2 GDPR FINES BY COUNTRY (TOP 10 IN VALUE)

Country	Total Fine value (EUR)	# Fines
UNITED KINGDOM	314,990,200.00	2
FRANCE	51,100,000.00	5
GERMANY	24,619,925.00	16
AUSTRIA	18,070,100.00	8
BULGARIA	3,173,370.00	9
THE NETHERLANDS	1,360,000.00	2
SPAIN	1,179,600.00	31
POLAND	933,868.00	5
GREECE	550,000.00	3
ROMANIA	445,000.00	12

COMPLEXITY BY DESIGN

Data collected for different usage, like social media have a completely different level of storage and security needs compared to data collected for something like a sales contract with an individual for a mobile phone. Similarly, the volumes of data are also different on a social media website which has over million messages per day compared to a retail store which processes thousands of invoices per day.

A. HOW MUCH 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 photos stored in free on a google drive with 15GB capacity. 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, which you may know how much it is. 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 within society and stored with repeated copies make this huge data set.

B. WHERE IS THE DATA STORED

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 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 more commonplace. Fig 5-2 below from the IDC DATA Age 2025 represents the shifting trend of data storage.

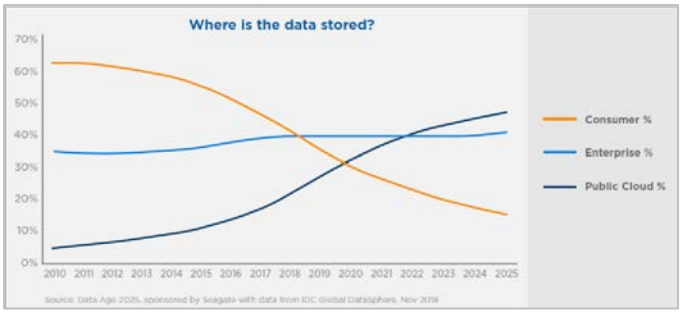

Figure 5-2 Where the data is stored (trend and forecast)

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.

C. DIFFERENCE IN DATA

Data is stored in different forms for different usage. In terms of its storage location, structure and source, data can behave differently, when it comes to revealing or hiding personal information. For example, a comment in a social media network like Facebook may not be revealing anything about the person posting the comment, though it may be specifying the name of the one who posts the message. Though a feedback given for a hotel stay, maybe stored in a database, in a table, along with full information about the person and details of his/her stay in the hotel. Similarly, for a sales contract, the company making the sales of suppose a mobile phone, may be storing the contact details of the person purchasing the phone, his/her address, and many other information in one place related to that particular transaction.

D. ENTERPRISE 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 be revealing 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.

E. STORAGE TECHNOLOGY

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. And also, 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 grows very high in volume and footprint across the globe.

F. 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 tabular form. This dataset is stored in data bases 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, but very significant in the world of mobile technology.

G. SPECIAL CHALLENGE WITH GRAPH DATASETS

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 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 presented for an unknown person, it is very easy to associate the same and find out who is the individual, revealing an individual's movement across different locations, by identifying a pattern[10].

H. WHAT IS CURRENTLY IN DEMAND

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 has more attention than the Enterprise data area from the technologists and researchers. As the boundaries between Enterprise data and nonenterprise 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.

HOW WE PROTECT DATA PRIVACY – WHAT IS ANONYMIZATION

Anonymization is primarily defined as a process for information sanitization whose intent is privacy protection. Different industries see it differently. For example, medical information about a patient is recorded in public health records[11]. 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 may be not. It may so happen that the patient attracts some unnecessary attention, which he/she is not seeking and may be uncomfortable with sharing such information.

A. WHAT NEEDS TO BE ANONYMIZED

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 color of hair of a person if combined with his address, may be used to identify who the person is. So the color 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 purposed but deleted after its use for an intended purpose, leaving no trace to identify the individual.

B. HOW IT WORKS

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.

C. 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.

1) 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.

2) 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.

3) 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 be revealing any information to trace back the individual and find out specifically, whose data is being presented.

D. DIFFERENTIAL PRIVACY WITH 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 [12] or another form or 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 regenerated 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.

E. WHAT'S ACHIEVED

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.

TECHNOLOGY – IS IT ENOUGH

In a paper submitted in Open Identity Summit in Bonn, 2019, authors argue that "Anonymization is dead, Long live Privacy". They advocate a paradigm shift, away from anonymization towards transparency, accountability and intervenability [13]. 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. There is a constant research also for algorithms to brake anonymity and de-anonymize information, proving that the methods are not sufficient in general and even methods considered sufficient, applied to some data, have been proved to be insufficient.

A. CONTEXT TO DATA

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 identify 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 method of making the data not identifiable to personal level, it is very difficult to classify itself, which data can be personally identifiable, and context to data is very important, which is most of the time, not stored with data itself.

B. ENTERPRISE DATA CONTEXT

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.

C. VALUE DRIVING MISUSE

Though this data about the customer can be misused too. 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 reaching the individual and manipulating his 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.

WHO BEAKS THE TECHNOLOGY – PROFIT SEEKERS OR MANIACS

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 unexpected anonymized data set has been used to reveal personal information.

A. 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.

1) 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<ref8>. 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, 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, deanonymized. This can adversely affect the individual.

2) 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.

3) 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.

4) 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 from this access of organization records about the employee. Also, they may gain access to the same information that the employee has in the that organization [15].

B. MOTIVE OF ATTACK

Motive of attack is am important consideration in preventing attacks as in case of any crime. The motive of attack, clearly hints the incentives to prevent attack. If the attack is done for profits, its for benefit of an individual not a society, but if the attack is done for no profits, who is behind the attack and what is the real motive, needs to be identified as it may be for a good reason.

1) PROFITS

With many of these 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 favors 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.

2) NOT FOR PROFITS

Apart from these direct attacks, which are for immediate benefit by the attacker, there are attacks, which were done for no profit seeking 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 saves other attacks and misuse of data and help the practice of data security advance further.

WAKE UP TECHNOLOGY – CAN TECHNOLOGY BEAT TECHNOLOGY ALWAYS

Technology is created based on certain theories and research. Hence the attacks to break technology as well. There are several methods, created with accuracy, which cannot be broken down by attackers[18]. Things like hash algorithm, which were supposed to be unbreakable, could be broken with advanced power of computing, but new hash algorithms were created with more sophisticated technology and are not broken yet, like the HSA256. Similarly, the field of cryptography has seen a very successful history, with evolving practices and technology to safeguard data and its subjects. The researchers are ahead or the hackers, is always a question of subjective bias, as who is a researcher and who is a hacker. Though considering the geopolitical situation across the globe, it has been quite established fact that hackers are not different than the creators of technology. Something which is loss for one is seen as an opportunity by the other.

A. CASE WITH NO PERSONAL DATA

Netflix, a popular movie streaming service, published the comments on movies removing the individuals who have made those comments. Netflix assumed that by removing the individuals, there is no conflict with privacy. But in due course of time, researchers proved that with only little knowledge of the users from some other movie

service, it was easy to identify the people who gave those comments on Netflix by Cross reference attack[19]. Do we call such an attack as an attack by Maniacs or by profit seekers?

Actually, this attack was done in 2006 and the attackers were not profit seekers but researchers. This attack prevented Netflix from starting their second similar challenge. In one way, it was an eye opener at its time and gave one of the first alarms on how sensitive data is and what seemed to be impossible can be easily possible.

B. CASE WITH UNIQUE ATTRIBUTES – SPARSE DATA

In a famous case in US, data of the Governor of Massachusetts [20], could be identified, with matching the Date of Birth and zip code of address for the person, against other identifiers for Mr. Governor like his filing for election etc. Just with this little information, it could be said with certainty that the person was Mr. Governor. As zip code and date of birth are not common. In a given area, with several people living, rarely two people are born on the same date. So just getting the zip code of the area and the date of birth of the person in one place can be misused to identify the person.

With different patterns for zipcode across the globe, some places, zip code is called, pin code, some places EIR code and some places with other name, and also the way zip codes are organized is quite different. For example, in Ireland, every individual house has its own EIR code. In UK, a small location has a given EIR Code, and in a location like India, a pretty large population has one PIN code. Hence this rule to match a date of birth with an address based code, may not be leading to a unique person in a data set. And there may be exceptions, but important point to note is that this can be as option to expose personal information. This is good enough to create safeguards, to avoid such an exposure. Can we say in this case, it's the fault of technology, or maybe on a lighter note, the fault is for the state which allocates same address codes to more than one people born on the same date?

C. TECHNOLOGY – IS IT SUFFICIENT TO ENSURE DATA PRIVACY

What we advocate is that technology alone may not be enough, to provide a completely safe data security framework [21]. There is a human element, which drives the need of technology. There is a human element, which also drives the need of hacking this technology (so that hackers/ sole profit seekers cannot harm individuals). So, what is the resort to provide data security to individuals. Based on the discussion for a social drive turning into legal drive, it can be easily seen that society is definitely looking to provide more security to individuals, so that innovation can flourish. Vulnerable individuals are provided protection from the resource rich parties, in order to avoid any temptation of being tricked into giving up whatever they have in hope of leading a better life. Even then the cost they pay to better their lives is at times compromising their integrity.

D. TECHNOLOGY FOR TECHNOLOGY OR HUMANITY

Anonymization technology is designed in order to prevent and data privacy loss. But at the same time, its technology and has, as in most other cases, more than one purpose. Not only it has to protect data privacy, but also it has to retain usage of data after anonymization. For example, if all names and all personally identifiable information is removed from a dataset, like Date of birth, address, card number, any unique transaction value, transaction timestamp, and so on, the dataset, may be of no use as these would be needed to make any meaningful analysis of the interaction with this individual by the institution, which is storing this data and intends to analyse it. This make the anonymization technologies, vulnerable in one way or the other, as they have to ensure that the post anonymization utility of data is retained. This may clearly mean some sort of compromise for the method to be chosen. This itself may lead to attacks and may need additional steps to improve the effectiveness of method of anonymization [22] in order to protect data privacy in view of this vulnerability.

E. THE TIME FACTOR

If technology can always beat technology be true, it's the time lag, which it takes between the creation and the breaking of a technology, which is the lifecycle of that technology. There has to be identified frameworks, which advocate the use and also evolution of technology in order to protect the privacy of individuals. Of course, putting

legal directives can be one of the methods to keep check on what can be done by the resource rich parties, but at the same time, these guards have to be flexible enough, that the individuals can enjoy the benefits of technology.

What comes out to be important is not possibly a static framework but an evolving framework, which enables to keep in line with the latest technology and also enables the new innovations using these technologies.

WHY NOT PROFITS – WHO HATES MONEY

Entities storing information on individuals, often have great utility of this data, in order to drive efficient handing of these individuals as their customers. Also this vast amount of data they store can be used by these parties, in doing machine learning and other innovative solutions, which can drive efficiency and cost saving for these companies. The analytics on the dataset apart from the analytics needed for driving operations, may also be a source of income for these companies. Companies can provide many value added services by analysing the data of a group of individuals and using the conclusion from the wide data analysis to provide one individual with a value added service.

A. INTERNET FACILITATES AND CREATES VALUE

An example from data on internet can be that small information like what an individual is searching on a search engine, itself is good enough for a website to publish a relevant advertisement and make money from the advertisement. If a company can identify what triggers a user to buy a product by analysing their social media feed and their behaviour on internet, it can be a great deal for the company to increase their revenues from their customers. If an electricity supply company can predict the consumption of electricity of its customers, it would be in a very good position to save on its electricity purchase from the grid, saving it a lot of money. If a company can predict with a good accuracy the price of a product and the optimal volume which users would be happy to buy, the company can magically improve its margins.

B. WHAT IS THE URGENCY

Not that these things have never been done, these are done by companies all through these years. What is added with the new technologies – to handle big data, communicate in real-time with 5G, store massive amounts of data and retrieve with unlimited calculation potential with cloud computing, and compute month long calculations within seconds using quantum computing (the list goes on) – is a great potential to do more. Provide more value added services to individuals and at the same time reward the innovators, to promote innovation based societies.

C. WHAT TO DO BY PRIVATE PARTIES

'Other Parties' have to carefully select the methods to anonymize in order to ensure that the objective they intend to use data for, do not in any way affect any specific individual whose data is concerned. Also, in no way, there should be a possibility of identifying the individual at the end of analysis or via other methods of attack on anonymized data, to protect the individual. Once this has been achieved, the individual is no more vulnerable.

D. CONTRACTS INSTEAD OF CONSENT

This may or may not be possible for data in a public domain, though for data in a private domain, where the access to data is restricted, this can be definitely possible. Those who have access to such data, are bound by confidentiality agreements on data protection and use for only a specific purpose to perform their duties. In case of Enterprise data, it is common to have the data access pretty bound by the contractual agreements between the people who have access to data and the organization which holds ownership to that data. If an organization intends to use the information it has collected from individuals, for a purpose, for which the information has not been collected, the organization is restricted by laws in doing so. Moreover, if the purpose for which the organization wants to use this data is covered in the contractual agreements between the organization and the individual/ party, whose data is in question, then still, the access of this data has to be ascertained in such a way that there can be

no harm/ disadvantage for the individual. Anonymization can easily provide the needed alteration in data in order to protect the individual in such cases.

E. WHAT REMAINS AFTER ANONYMIZATION

After this anonymization, if the data is suitable for any use, is definitely not the case. As the anonymization would alter the utility of data for post anonymization usage. For example, if the data is stripped of personal information like contact number, there is no way after anonymization to contact the individual, hence in this case, the data cannot be used as a basis of running a marketing campaign via phone. Though this data can still be used to find out other important facts about the marketing campaign. Segmenting the users, with random ids, can help the company find out what attributes are there for running a successful marketing campaign, based on the purchase history of these users, whose identity has been replaced with these random ids. Many other usage of this data are possible like in case of fraud detection, demand forecasting and likes.

F. WHO GETS THE REWARD

Innovation always comes with new usage and expectations never seen before. And the innovator in most cases expects a

reward for his/her innovation. The rewards can be in any form and may not be monetary. Innovations in general do not consider rules and regulations like the ones for data privacy. Innovations lead to new rules being formed in order to make the innovation benefit individuals in particular and society in general.

G. IS DECISION SCIENCE WORTH A SPECIAL MENTION FOR DATA SECURITY

With new set of tools in the recently prominent field of decision science also known as data science, there has been several new methods invented with the aim of discovering new information from data. Though business users and Enterprise has traditionally relied on data and experience to take informed decisions, decision science tools, have been quite instrumental in making this practice more and more scientific and transparent giving it a structure and repeatability, also substituting “experience” in some sense. The practice of creating machine learning models for a machine to learn from past data, is practically an attempt to replace experience in the traditional decision-making process.

Though this is still an emerging field of study and yet to find an acceptable place in day to day business of the Enterprise, this practice has created a high demand for data to be available for machines to learn. Such data has never been exposed to any individual who would design and review and prepare such machine learning algorithms. This has created a void, where technology is ready with a lot more to offer. But businesses cannot spare so much of data to this technology, due to restriction on data exchange and high risk in this valuable (including personal data) being misused. Anonymization can be a solution to fill this void and open new possibilities.

H. WHO MAKES MONEY

Companies which are leading in the field of machine learning are definitely leaving no stone un-turned to make it reach the masses. For example, the use of location data from individuals to provide accurate travel statistics to other in general using differential privacy to protect data privacy while still using this data generated by individuals [23]. Another example of Apple using differential privacy in ios10 to improve quick type and emoji suggestions and others [24] [25]. These companies have proved that the usage of anonymization techniques can definitely open new use cases and can benefit the individuals with innovations, at the same time ensuring data privacy.

I. CHANGING PARADIGM IN SOFTWARE WITH CLOUD ENVIRONMENTS

With the data increasingly residing in cloud environments, the importance of data security is further amplified. Most of the users of social media technology are not even aware of the fact that their data is available in public

cloud environment. Though this does not pose any risk, but the recent breaches due to failed security standards maintained by companies on MongoDB [26] leading to compromise of several accounts clearly indicates a need of awareness.

Not only social media, but all the Enterprise companies, colleges and eventually everyone including governments as well, may move to cloud environments to store their data. All this not alone for technology adoption, but definitely for saving money too. Technology and innovation has always helped companies to save money. Money has always been a driver for innovation. Hence we cannot ignore the fact that data privacy can be taken as an excuse to restrict "other parties" from using the vast store of data they have for deriving more workable information, in order to benefit the individuals at the same time, making these other parties increase their profits and reward the innovators.

CONCLUSIONS

In this paper, we have analyzed the different aspects related to Data privacy in addition to the technology aspects. As data is never alone, but always has a meaning with a context, so is the technology for data security. The technology used to provide protection to the privacy of individuals, needs to be safeguarded with appropriate framework, in order to evolve itself. GDPR and other such framework being established form a good basis on data management. 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.

For public data availability, which is a primary responsibility of Govt. and other public entities, it is vital that not only the personal information, but also personally identifiable information is appropriately redacted or anonymized appropriately from the primary collected information. There is a constant need to keep a watch on the technology getting outdated. It is of paramount importance to keep updated with the latest technology in case of data anonymization. Would this also advocate removing some publicly available datasets, from time to time – maybe not, but for sure, keeping a close watch on the publicly available datasets and remove them, as soon as any vulnerability is reported by the researchers or the hackers. A proactive approach would always be the best under a defined framework.

When these OTHER parties, want to utilize the individual's data to improve their services in general and not specific to that particular individual, they need to anonymize the data. The 'other parties' cannot use the data from individuals for any purpose other than what the data was collected for, as per the data protection regulations like GDPR, CCPA and others across the globe. This clearly indicates that the companies and other entities storing information on individuals, must have proper safeguards in place to protect the privacy of individuals. 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.

[2] "2019 Data Breaches: 4 Billion Records Breached So Far." Norton, Security centre, emerging threats.

[3] J. Wright, "Teen blasted out of Rod Laver after 'Timebomb' tweet," The Sydney Morning Herald, 08-Jul-2013.

[4] X. Qiang, "The Road to Digital Unfreedom: President Xi's Surveillance State," J. Democr., vol. 30, no. 1, pp. 53–67, Jan. 2019.

[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.

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

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

[8] W. Vogels, "Eventually Consistent," Commun ACM, vol. 52, no. 1, pp. 40–44, Jan. 2009.

[9] S. Gambs, M.-O. Killijian, and M. Núñez del Prado Cortez, "Deanonymization attack on geolocated data," J. Comput. Syst. Sci., vol. 80, no. 8, pp. 1597–1614, Dec. 2014.

[10] L. Sweeney, Datafly: a system for providing anonymity in medical data. Database Security, XI: Status and Prospects, T. Lin and S. Qian (eds), Elsevier Science, Amsterdam, 1998.
[11] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L.
[12] Sweeney, "Privacy Preserving Synthetic Data Release Using Deep Learning," in Machine Learning and Knowledge Discovery in Databases, Cham, 2019, pp. 510–526.
[13] J. Zibuschka, S. Kurowski, H. Roßnagel, C. H. Schmuck, and C. Zimmermann, Anonymization Is Dead – Long Live Privacy. Gesellschaft für Informatik, Bonn, 2019.
[14] J. Krumm, "Inference Attacks on Location Tracks," in Pervasive Computing, Berlin, Heidelberg, 2007, pp. 127–143.
[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," Int. J. Comput. Netw. Inf. Secur., vol. 9, no. 1, pp. 1–11, Jan. 2017.
[16] Mohurle, Savita; Patil, Manisha, "A brief study of Wannacry Threat: Ransomware Attack 2017 - ProQuest." International Journal of Advanced Research in Computer Science; Udaipur Vol. 8, Iss. 5, May 2017.
[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] F. Liu and T. Xie, "How to Break EAP-MD5," in Information Security Theory and Practice. Security, Privacy and Trust in Computing Systems and Ambient Intelligent Ecosystems, Berlin, Heidelberg, 2012, pp. 49– 57.
[19] A. Narayanan and V. Shmatikov, "How To Break Anonymity of the Netflix Prize Dataset," arXiv:cs/0610105, Nov. 2007.
[20] D. Barth-Jones, "The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2076397, Jul. 2012.
[21] L. Sweeney, "Weaving Technology and Policy Together to Maintain Confidentiality," J. Law. Med. Ethics, vol. 25, no. 2–3, pp. 98–110, Jun. 1997.
[22] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei, "Anonymizationbased attacks in privacy-preserving data publishing," ACM Trans. Database Syst., vol. 34, no. 2, pp. 1–46, Jun. 2009.
[23] "Tackling Urban Mobility with Technology," Google Europe Blog.
[24] "Apple Previews iOS 10, the Biggest iOS Release Ever," Apple Newsroom.
[25] H. B. Kartal, X. Liu, and X.-B. Li, "Differential Privacy for the Vast Majority," ACM Trans Manage Inf Syst, vol. 10, no. 2, pp. 8:1–8:15, Jul. 2019.
[26] DK Kola, S Barre, S Medipally, A Gaikwad, "Information governance failure in MongoDB," From the selective works of Dinesh Kumar Kola, University of Cumerlands, September, 2019.