Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 5, Issue. 1, January 2016, pg.334 – 340

Digital Traces: A Survey On Whether Are We Being Tagged Vendible Without Our Consent

Mrs. Uma.N[1], Mrs. Swetha.G[2]

[1]Department of CSE, New Horizon College of Engineering, India

[2]Department of ISE, MVJ College of Engineering, India

[1] umamam@gmail.com; [2] swethalahari@gmail.com

Abstract— In the era of information technology, due to different electronic, information and computer technology devices and processes like cloud, social networks, Internet activities, the amount of data that is being created and stored is almost inconceivable and it is just growing exponentially. According to Pat Gelsinger, CEO of Vmware and former CEO of EMC corporation, "Data is the new science and big data holds the answers". Big data is a term that describes the data which are huge in volume, veracity, velocity and volatility. In big data analysis, we have to collect data from different sources, and thus have to deal with structural and non-structural data. The data can be collected from social networks, our day to day digital traces we leave like credit card transactions etc. Thus there arises a question on personal data and privacy protection. To handle the big challenges of privacy and security in big data analysis, the responsibility of the personal data in accordance with user preference should rest with the provider than with the user. But with increasing apps, which we download from large app stores, the intermediaries can create a market place where there can be a negotiation of community standards for privacy. In this paper, we try to analyze whether the personal data which are stored and analyzed can become a commodity without our consent. Also we try to explore over this challenge and the different implementation privacy approaches.

Keywords— Big data, Data Brokers, Data Privacy

I. INTRODUCTION: WHAT IS BIG DATA

With the rapid growth in the development of electronics and computer technology, the amount of data generated is huge in volume, velocity, veracity and volatility. We use the term "Big data" to represent such data. Big data analysis is used by many organizations to turn data into insights and to turn insights into revenue.

II. BIG DATA TECHNOLOGY

A) The process Pipeline

When we are considering the Big data pipeline, there are mainly 5 stages-Data Acquisition, Data Extraction, Data Integration, Analysis or modelling and finally interpretation of results.

Data Acquisition: Data acquisition can happen from various sources like survey forms, credit card transactions, navigation sites like google, yahoo. Data can also be collected from the various app permissions, social networking sites like facebook, twitter etc.

The challenge to deal with in Big data analysis is the variety, volume and velocity of data. With privacy, the biggest challenge is that some of the data requires permission from third parties where as some does not. Many of the free services that we avail like Gmail, facebook, etc ,we believe that we "own" the data ,but in reality is maintained by a third party service provider. Information can also be collected from credit card companies, life insurance industry, Medical and Lab records, etc in a form that are computer understandable and robotically resolvable.

Extraction: Once the data is collected, companies use various tools to extract the data they are looking for and check for the correctness of the data.

Data Integration: Data integration is the next process in the pipeline where the data from various sources are represented in a form that is computer understandable and robotically resolvable.

Data Modelling: In this process stage, various conclusions will be inferred based on the data collected by analyzing using various models like predictive models, statistical models etc. Various hidden patterns, frequent patterns, fluctuations etc are analyzed and lots of inferences can be made. For example, Chase Bank used predictive models to predict risk of foreclosure on their mortgages.

Data Interpretation: In this stage, all the assumptions are re-examined and the errors in multiple stages are found out. Also the entire analysis will be retraced to derive the useful insights.

B) Where is Big data used?

Big data is used by various world famous organizations like Google, Yahoo, Amazon, Facebook, and Netflix for improving their B2B(Business to Business) interactions.

Google is definitely the pioneer and leader regarding the quantity of data it has to process. They have developed "Map reduce Algorithm which helps in their analysis method. Similarly, Yahoo has developed the software Hadoop, a java oriented open framework which is used for processing big blocks of data in distributed system.

Netflix is an American company which provides video content on demand. Based on their predictive analysis, they could even predict which movies and series will be accepted by the users and also the standard acting qualities expected from actors.

These companies use these data to find the needs and preferences of their customers for the purpose of improving their services. But apart from these experts, there are marketing agencies which do probability analysis on the data and give recommendations to companies regarding the choice of products preferred by customers.

With the improvement in Internet of things technology, we have various downloadable apps which monitor even the microlevel daily activities like calorie intake, workout time and intensity, sleeping rhythm etc.

Banks use Big data technology to predict and prevent credit card abuses and to analyze their clients behaviour, their needs and to create personalized offers through their products.

Big data is used in health care system [1]. Big data information is used to recognize [2] people with a high risk of certain medical conditions at early stage and providing improved quality care and lowering the increase cost of health care.

Moreover, social networks act as the biggest source of data. By analyzing the activities and status on social networks, Google could have accurately predicted the course of swine flu epidemic spread by monitoring the swine flu related symptoms search on the internet in 2009 before there was a major news coverage [3].

National Security Agency (NSA) uses the big data analysis to uncover security threats and prevent terrorists attacks. Big data technology is benefiting individuals either directly or indirectly in all walks of their life[4].

C)Big data mining : A friend or Foe : Crisis management and big data analysis shows that use of personal data without owners consent is crucial in rescuing operations, but not when it comes to commercial and marketing purposes.

Case Study 1: One of the major applications of big data is marketing where the marketers try to place their products and services in front of highly targeted customers. But, when the customer is categorized into one category based on their behaviours, there is possibility for harm. In spite of the possibility for harm, marketers still use big data to aim at people on social media platforms like search engines and email. Forceful entry into personal area by providing advertisements based on friends, likes and email content is causing anxiety among consumers. Big data analysis proved to be scary, when a statistical model deployed by the guest marketing analytics at Target [5], an American company accurately predicted the pregnancy of a teenage girl based on the statistical study of her buying patterns even before her family knew. Spooky to contemplate, living in a world where Google and Facebook and your phone, and many companies like Target know more about you than your parents do.

Case study 2: Google faces a law suit [6] in California over whether bulk scanning of emails to deliver advertisements breaches state and federal wiretap laws. In its filings for the lawsuit, the company has also admitted scanning the contents of emails sent and received by American students who attend schools which use the company's Apps for Education suite. This actually raises a serious question of googles data mining the emails of school going children for their financial gain.

Case Study 3: Mortons steak house: After a customer jokingly tweeted a Chicago based steak house chain ,Morton's steak house expressing his wish to have dinner after his long, tiring day and that he would land in New York airport[7],the team at steak house saw the tweet, responded by sending a delivery man at the New York airport with his usual order pulled out from their database.

III) Does Big data know about us?

From the above mentioned few of the case studies, we can draw a conclusion that data collected and processed using Big Data Technology can easily reveal personal identities of an individual.

Current Research shows that issues of privacy concerns on the internet, 84% of the respondents affirmed on their privacy concerns. The Pew research centre recently published a new privacy poll on Americans views about data collection and security. Survey report shows that about 74% [8] American population feel insecure about the invasion of their privacy. To appease such privacy concerns, the organizations using Big data analysis claim that it is impossible to link records with actual persons who left traces. In the article titled "10 big data privacy problems, Rebecca Herold, gives several examples which proves that anonymization is almost impossible with even powerful analytics.

Information such as user names, e-mail address, web pages visited, inquiries done, comments left, IP addresses, times of action and many, many more information are sent through web browsers. We leave traces. Regardless of that, individuals voluntarily give up their privacy in order to participate in all the services provided on the Internet. It is very easy to monitor the location and movement of persons via their smart mobile phones, however such knowledge does not diminish the number of smart mobile phone users.

Data brokers [9] collect and maintain data on hundreds of millions of consumers, which they analyze, package, and sell generally without consumer permission or input. Since consumers generally do not directly interact with data brokers, they have no means of knowing the extent and nature of information that data brokers collect about them and share with others for their own financial gain.

Data brokers collect and sell information for a variety of purposes including for fraud prevention, credit risk assessment, and marketing. Their customer base encompasses virtually all major industry sectors in the country in addition to many individual small businesses. Some of the most well-known products sold by data brokers are credit reports that businesses use to make eligibility determinations for, among other things, credit, insurance, and employment – activities where consumers have detailed statutory consumer protections regarding the accuracy and sale of their information.

IV) Challenges in Big Data Privacy

Big Data privacy can be defined as the prevention of disclosure of all the PII (Personal Identification Information) and other sensitive information collected as a part of Big data Acquisition. As we have analyzed privacy preserving is a big challenge in big data. The various factors which play an important role in big data privacy[10] and make it more challenging are as follows:-

A) Context based Privacy:

The most basic privacy challenge is context based privacy definition in which each data set have different meaning in different context and deciding which data set is sensitive in that context is very difficult. For example , a data collected say 31 ,can mean many things-Date, Dollar, Days, Birth date, Age etc.indicating how much privacy will be required for different data set in different context becomes even more difficult and thus applying such privacy becomes challenging.

B) Co-related and aggregated data sets:

The various data sets that we collect in acquisition are related to each other. Useful conclusions can be drawn by analysing the relationships among data sets. Hence, if there is a discloser in privacy of one data set ,it may lead to privacy discloser of other data set. Privacy threat may also occur at the time of data processing since there is chance that one information of one data set maybe required by the other. Such co-related and aggregated data set are conceptualized as Quasi-identifier which is a big threat to big data privacy.

C) Threat Modelling

Thread modelling is a systematic & structured technique of designing a privacy preserving solution by prior identification of privacy objectives and attacks that might occur. So such an approach requires a crystal clear idea about what type of privacy threat maybe possible and how to handle them. This is one of the biggest challenges in big data which is huge in volume, variety, velocity etc.

D) Privacy Budgeting

The cost that should be spent to preserve privacy in big data is another very challenging issue. The cost estimation is done in terms of computational requirements, thereby restricting us to choose techniques that are computationally very expensive. Thus providing good utility and privacy in lower budget could be possible only if we have efficient computation and that is even more challenging.

E) Policy and Legal ramifications

Today data is treated as a valuable asset and due to its ever increasing importance, the policy and legality plays very important role. Different countries have different policies and laws for data and its privacy. A new technique is required which should follow all the diverse aspects of the law. Preserving privacy with fulfilment of all this legal constraint is a big hurdle.

V) Approaches for ensuring privacy with big data technology

Big data privacy can be ensured using two approaches: privacy by design and privacy by imposing rules and regulations. But definitely privacy can only be achieved only by developing an approach which is capable of handling both technical and legal aspects. These two approaches should go hand in hand. Among the various approaches to ensure Privacy by design, the most popular ones are given below.

A) Encryption based

In this approach, we use cryptography based encryption and decryption techniques on the data sets. The main problem with this approach is that cryptographic techniques are computationally intensive and thus degrades system speed and increases the cost. More degradation in performance can occur with increase in datasets.

B) Anonymization based

Privacy preservation approaches like K-anonymity and other anonymization techniques are based on anonymization principles. In this approach also, it is possible to derive sensitive information from database without knowing precisely which records belongs to target individual. As an example let us suppose that we are looking for a man who is 6 foot tall living on a particular block, and if the database shows that all of them are having some medical issue like hypertension, then the adversary would be able to get this additional information

about the target, even without knowing who is the target individual-anonymity doesn't take in to account the background knowledge of information from other sources. It is also vulnerable to active attacks since there are cases where adversary may know some sensitive information and could compute the generalization schema from data and may reveal more sensitive information's from the database.

C) Noise based approach

In this approach the noise is added in data set. In these databases, privacy is obtained by perturbing the true answer to a database query by the addition of a small amount of Gaussian or exponentially distributed random noise. The results for noise generation are of independent interest.

D) Differential privacy

It a privacy approach in which probability of output of two different data set will nearly be same. When two different data set produces nearly same output then the adversary can't determine the actual targeted data set by any quasi identifier. Differential privacy is very suitable for big data. It is not computationally intensive like encryption based approaches and at the same time addresses the problems of k- anonymity. It makes it so hard for the adversary to infer the presence or absence of any individual, even if the adversary knows the exact information of all the remaining individuals in the data set [11].

V) How can one protect privacy in the era of big data

As an individual, here are some means through which we can protect privacy.

- **Quit sharing so much on social media:** If we have a very limited crowd for which we need to share some media, it is advisable to share it individually.
- **Don't provide too much information to businesses or other organizations :**It is advisable not to share additional information with any organization, other than for the purposes for which you're doing business with them. Unless required it is unnecessary to provide out PII (Personal identification information) like name, phone number, address etc.
- **Use an anonymous browser:** It is advisable to use anonymous browsers like Hotspot Shield or Tor (The Onion Router) when visiting sites that might yield information that could cause people to draw inaccurate conclusions about you.
- **Create awareness about privacy issues and educate the contact list in social networks:** Ask others not to share information online about you without your knowledge.
- **Turn on cookie notices in your Web browser, and/or use cookie management software or infomediaries:**

"Cookies" are tidbits of information that Web sites store on the computer, temporarily or more-or-less permanently. They may be passwords and user Ids, so that we do not have to keep retyping them every time we load a new page at the site that issued the cookie. There are cookies which can track our motions through website ,the amount of time we spent there, the links that we click on and so many such information which the organization looks for their marketing purposes. Most cookies can only be read by the party that created them.

They can track which pages you load, which ads you click on, etc.***, and*** share this information with ***their entire*** client Web sites

VI) Conclusion

Today, we are digitally connected, socially networked and better informed. Customers live their lives "in the moment," updating their relationship status, interacting with their friends and sharing their likes, dislikes and opinions, all in real time through the power of their mobile devices. They are literally changing the rules of engagement and, through that, becoming more empowered. It's easy to understand in the context of business-to-consumer (B2C) interactions and how much big data has benefited in many realms of life. Consumers are conducting more and more of their daily business online and through their mobile devices. With these activities, consumers are creating a voluminous and unprecedented trail of data regarding who they are, where they live, and what they own.

Today, a wide range of companies known as "data brokers" that largely operates hidden from consumer view, collect and maintain data on hundreds of millions of consumers, which they analyze, package, and sell generally without consumer permission or input for their own financial gain.

But such massive data acquisition is clearly a threat to privacy. Laws do exist, however, regulations and provisions are not able to protect privacy and at the same time enable the free use of Big Data technology. As an individual, one should be aware of such privacy threats and be more responsible with what they do with the data, rather than completely expecting the government and legislations to reinforce the control over the data. Clearly privacy by design and privacy by regulations should go hand in hand. In this paper, an effort is made to know more about the buzz word "Big Data Privacy" and also to understand the challenges incurred in ensuring the same. Also an effort is made to understand the various approaches used and a comparative study is made on them.

REFERENCES

[1] **"**Big Data is the Future of Healthcare", Cognizant 20-20 insights, September 2012.

[2] "Big Data Analytics" ericsson White paper,284 23-3211 Uen, August 2013

[3] Google could have caught swine Flu Early http://www.wired.com/2009/04/google-could-have-caught-swine-flu-early

[4] NSA mass survellience-Biggest big data story http://www.forbes.com/sites/metabrown/2015/08/27/nsa-mass-surveillance-biggest-big-data-story/

[5] Big Data:What is even more creepier than Target guessing that you are pregnant http://www.slate.com/blogs/how_not_to_be_wrong/2014/06/09/big_data_what_s_even_creepier_than_target_guessing_that_you_re_pregnant.html

[6] Google faces law suit over email scanning and student data ***http://www.theguardian.com/technology/2014/mar/19/google-lawsuit-email-scanning-student-data-apps-education***

[7] The greatest customer service story ever told-starring Mortons steak house http://shankman.com/the-best-customer-service-story-ever-told-starring-mortons-steakhouse/

[8] greatest customer service story ever told-starring Morton's steak house http://shankman.com/the-best-customer-service-story-ever-told-starring-mortons-steakhouse/

[9] Americans view about data collection and security. ***http://www.pewinternet.org/2015/05/20/americans-views-about-data-collection-and-security/***

[10] T.Krishna Mohan Pd Shrivastva , M A Rizvi , Shailendra Singh "Big Data Privacy Based On Differential Privacy a Hope for Big Data" , in 2014 Sixth International Conference on Computational Intelligence and Communication Networks

[11] Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating noise to sensitivity in private data analysis." In Theory of Cryptography, pp. 265-284. Springer Berlin Heidelberg, 2006.

