Mohammad Hossein Ariana *et al*, International Journal of Computer Science and Mobile Applications,

Vol.4 Issue. 4, April- 2016, pg. 11-19 **ISSN: 2321-8363**

Impact Factor: 4.123

Big Data Era and Privacy Challenges

Mohammad Hossein Ariana[1], A Broumandnia[2]

[1]Islamic Azad University, Dehdasht Branch, Dehdasht, Iran, Ariana.h@iaudehdasht.ac.ir

[2]Islamic Azad University, South Tehran Branch, Tehran, Iran, broumandnia@gmail.com

Abstract

Today, we use most of time generating and sharing data using communication devices and social networks. Companies might collect and analyze these vast amounts of data, known as big data and put them in huge datasets. They might perform some processing on this data, i.e. big data analytics to get better insights about us and use them on processes like targeted advertising to improve their business and sometimes help us. But the point here is that out information might be mined for information about us and used against us. This process might violate our privacy. Different threats related to privacy are presented here and some solutions have been discussed.

Keywords: Big Data, Privacy Challenges, Anonymization, Analytics.

1- Introduction

Every day more smartphones and PCs are sold and more people get access to Internet and join social networks or use online services [1]. With the appearance of social network websites, users describe their lives by posting details of their activities, people they meet, places they travel, photos they take, and things they enjoy and like. As a result, organizations can access to large volumes of data through monitoring of customers activity, social networks mining, website tracking, sensors and other tools. This phenomenon is often termed Big Data [2].

Managing and extracting patterns from the produced data is a challenge and key to competitive advantage. Analytics solutions that mine this data are important as they can help organizations gain insights not only from their privately acquired data, but also from large amounts of data publicly available on the Web. The ability to cross relate private information on consumer preferences and products with the data extracted from tweets, blogs, email messages, product evaluations, and data from social networks opens a wide range of possibilities for organizations to understand the needs and priorities of their customers, predict their wants and demands, and optimize the use of resources. This paradigm is being popularly termed as Big Data Analytics.

Big Data present threats and opportunities for organizations. Those organizations are not only exploring different ways to analyze and exploit information hidden within it but also have to tackle with the cost and risk of storing that data. In the same manner, Big Data is a threat and opportunity for individuals, too. On the one hand, most individuals now have instant access to vast amounts of information, which provides a wide range of benefits, including opportunities for innovation, communication and freedom of expression. On the other hand, these new pools of data also include information about individuals, and the use of Big Data tools to combine and analyze that information could result in privacy violations.

2- Big Data

Big data is defined as the capture, managing, and analysis of data which cannot be handled using existing relational databases [3]. Big data usually involves semi-structured or unstructured files, emails, social network posts, likes, comments, videos, images, sensor data, log files, and any other data which is one does not find in usual database records [2].

Big Data is becoming related to almost all aspects of human activity from just recording events to research, design, production and digital services or products delivery, to the final consumer [4]. It can also be used in used in statistical machine translation, as suggested in [5], for example.

Big data has a set of characteristics, usually known as 5vs'. These characteristics are volume, velocity, variety, veracity and value and are expressed as follows:

Volume refers to the vast amounts of data generated every second. Just think of all the emails, Facebook posts, Twitter and Whatsapp messages, photos, video clips, sensor data etc. we produce and share every second. We are talking about Zetabytes or Brontobytes of data, not just terabytes. The overall created and copied data volume in the world was 1.8 Zetabytes in 2011 which increased by nearly nine times within five years [6]. This property makes datasets too large to store and analyze using traditional database technology, like relational database systems.

Velocity refers to the rate of generation of new data. Over two million petabytes of data are generated each day, and in last two years, more than 90% of the world's total stored data has been created **[1]**. Just think of social media messages going viral in seconds, the speed at which credit card transactions are checked for fraudulent activities, or the milliseconds it takes trading systems to analyze social media networks to pick up signals that trigger decisions to buy or sell shares. Big data technology allows us now to analyze the data while it is being generated, without ever putting it into databases.

Variety refers to the different types of data we can now use. In the past we used relational databases and hence our focus was only on structured data, such as personal or financial data (e.g. sales by product or region or name or address of customers). But today, 80% of the world’s data is now unstructured, and therefore can’t easily be put into tables (like photos, video sequences, email messages, sensor data or social media posts) [7]. With big data technology we can now harness differed types of data (structured, unstructured and semi-structured) including messages, social media conversations, photos, sensor data, video or voice recordings and bring them together with more traditional, structured data.

Veracity refers to the trustworthiness of the data, i.e. how much we can trust the big data and make decisions based on that. With many forms of big data, quality and accuracy are less controllable. How do we know that data is not spoofed, corrupted? [8] Big data and analytics technology now allows us to work with these types of data.

Value: The final v of big data value. It is good having access to big data but unless we can turn it into value. So you can safely argue that 'value' is the most important V of Big Data. It is important that businesses make a business case for any attempt to collect and leverage big data. It is so easy to fall into the buzz trap and embark on big data initiatives without a clear understanding of costs and benefits.

3- Privacy

First of all, let's define privacy and explore its meanings. We can define privacy as the right to be left alone. Privacy is the right of people to make personal decisions according to their own willing, it is the right of people to be free from things like unwarranted drug testing, illegal house entrance or electronic surveillance. Information privacy, which we discuss it, is the ability of an individual, group or a company to stop information about themselves from becoming known to people other than those they choose to give the information to.

Privacy and the Internet

Using the Internet can affect the privacy rights a person has in his/her personal data. Internet use generates a large amount of personal information which provides insights into your personality and interests. Privacy issues relating to identity include the possible use of a person's email identity and address. Ease of access to email addresses results in sending vast amounts of unwanted e-mails, known as spam which we are all familiar with. Identification through email and IP addresses and the ability to locate people's physical addresses easily through national and international directories have raised new privacy concerns.

Many of websites use cookies, which are small pieces of text, to better understand users' habits and send them back to their servers. Often, cookies are installed and used without the permission of user. These information may be used without our consent and even sold to third party companies, e.g. for targeted advertising.

Another threat to privacy is concerning the monitoring which is done in workplaces and universities. Many companies use CCTV cameras or special software and hardware systems like key-loggers to monitor their employees. They can find about the websites that employees visit, video and audio clips they share on social networking websites, e-mails they send, and the people they communicate with. This way, they can acquire a lot of information about their employees.

But why privacy is important? With dramatic technological revolution, as more data is being collected and shared, information privacy is becoming more complex every moment. Technology and techniques for big data analytics gets more and more sophisticated. These changes put organizations in an incredibly complex situation for ensuring that personal information is protected. Hence, privacy has emerged as one of the most significant consumer protection issues in the global information economy.

4- Privacy Challenges in Big Data

Until recently, privacy was not among the main concerns of users of computer systems. This vast variety of smart devices did not exist and network connections were poor. Popular services like Facebook, Twitter and Instagram were not born. With the growth of the technology, a set of new challenges appeared. Such challenges which one cannot propose a comprehensive solution nor ignore them. Big Data was one of these challenges which occupied the minds.

Big Data plays a main role with respect to privacy of the individuals. With the increased processing power of computer servers, a big concern about violating the privacy of users appears. Take for example, the targeted advertising process. This process states mainly that instead of sending the ads to all the people, we could have a classification of customers which are more likely to apply for our products and services. Most of us have probably seen ads in social networking websites and/or applications which are related to what we post or what we search. This job needs a platform for surveillance, storage, and analysis of users' behavior. Many of companies sell their users' data like search keywords, posts in social networks, e-mail messages and other statistics to advertising and third party companies. But should these companies stop using our data?

A point here to note is that different users have different opinions about privacy. Level of privacy might depend on the culture and customs of the people. For example, in some countries the income of people might not be a private but in some others it should kept private. As another example, some users might want to use targeted advertising and this process may help them find they want. Take for instance, the Amazon website when you buy a book and below that, there is a section named "people who bought this book, have also …" which may save you a lot of time searching for related books. So, electronic services may present us with a better experience using our own data.

Another point about privacy which has caused a great deal of controversy is regarding to the security and terrorist attack. Government agencies can monitor social networks and eavesdrop phone calls and e-mail messages and find about suspected people and prevent them from harming others. Shall we voluntarily sacrifice our privacy for increased security? This way, national agencies freely collect a lot of information about us and they might be used against of us.

So, one might not assume advertising companies and processing and storage systems as guilty. Sometimes the main problem is about the user awareness of the process. If the user knows that each time he/she stores data on a server and upload a video on Youtube or sends an email, this data gets in hands of other companies and his/her privacy is violated, he/she would not allow that service do this job.

So, big data does not necessarily violate individual's privacy. The problem is with the users being unaware of these technologies and misuse of service providers and finally lack of an effective and comprehensive mechanism to use this data which can threaten our privacy.

Privacy concerns must be considered in different phases of big data life cycle, i.e. collection, storage and processing [9]. Here we mention some of the risks that we might encounter by allowing our information intentionally or accidentally used by others.

a. Discrimination

The use of big data analytics by the public and private sector can now be used by the government and companies to make determinations about our opportunities to obtain a job, a license, or a credit card. When you apply for a job, they might look for your records they have collected from different sources and make a decision based on that. Actually, Big Data analytics provides the ability for discriminatory decisions to be made without the need for that explicit and obvious evidence, which can affect everything from employment to promotions, from obtaining a citizenship to fair housing and more.

b. Embarrassment because of data breaches

Every day, we hear that some attack has been done and a lot of information has leaked. Companies collect a lot of information about their customers and users. After data leakage, this data might be exposed and used by others to make harms to these customers and violate their privacy. Apart from attacks, the actions taken by businesses and other organizations as a result of big data analytics may breach the privacy of those involved, and lead to embarrassment and even lost jobs. Consider for example that a woman shares an intimate matter like pregnancy with some friend via email, and some company gets access to that information and using big data analysis predicts some personal details such as the due dates of pregnant shoppers and sends ads for her. In such cases subsequent marketing activities resulted in having members of the household discover a family member was pregnant before she had told anyone, resulting in an uncomfortable and damaging family situation. Even employers may decide to fire or demote employees with certain illnesses. Retailers, and other types of businesses, should not take actions that result in such situations.

c. Performing Re-identification

A big threat regarding the privacy in big is the ability to perform “re-identification”. Re-identification means that a big dataset in hand which has been anonymized to contain only unidentifiable data about individuals, is explicitly scanned for correlations that lead to a unique fingerprint of a single individual. More precisely, after linking different datasets together, one might find a lot of information about a certain individual and he/she might be identified, like identifying Governor William Weld in an anonymized medical database [10]. If data masking is not done effectively, big data analysis could easily reveal the actual individuals who data has been masked.

d. Government exemptions

Some countries have regulations concerning privacy of the people, like Privacy Act in USA. These regulations usually restrict people and organization from doing some processing on the big data. But often the governmental agencies including intelligent services are exempted explicitly or implicitly from these regulations. They might collect Personally Identifiable Information (PII) including name, any aliases, race, sex, date and place of birth, Social Security number, passport and driver’s license numbers, address, telephone numbers, photographs, fingerprints, financial information like bank accounts, employment and business information and more about each citizen and company. This information might be used against us. Even if we refuse to share information in social networks and read the privacy rules of services carefully, there are always projects like Echelon [11] and Prism [12] which might collect our personal data.

e. Accuracy of Information

Numerous companies collect and sell consumer information that is not clearly protected under current laws. There is also little or no accountability or even guarantees that the information is accurate. The data files used for big data analysis can often contain inaccurate data about individuals, use data models that are incorrect as they relate to particular individuals, or obtained from corrupted and bogus algorithms. Some companies might manipulate the data they want to sell for specific purposes, e.g. to mislead other companies to improve their own economic superiority or to affect competitors' business patterns. The companied might use this erroneous information about us and face us with unrelated and harmful proposals which might danger our health or economic situation.

Solutions

There are solutions presented for improving the privacy of Big Data, like aggregation and de-identification [9], anonymization, pseudonymization, encryption, key-coding, data sharding [13] , suppression, data swapping, adding random noise, synthetic data [10], k-anonymity [14] which are useful in some situations, but in general, there is no solution presented to be complete and guarantee our digital privacy.

5- Conclusions

We discussed here that Privacy in the right to be left alone and today, because of huge revolution in amount of data being generated and analyzed, known as big data, we have less control over our data and companied and government agencies might use our data against us. We have discussed some of these threats we may encounter. Some methods have been proposed but they do not always protect our full privacy. Before using modern technologies like Internet of Things in our daily life and putting our entire life online, more general and safe methods should be proposed and tested.

References

[1] Matthew S. Eastin, Nancy H. Brinson*, Alexandra Doorey, Gary Wilcox, 2015, "Living in a big data world: Predicting mobile commerce activity through privacy concerns", Computers in Human Behavior.

[2] Seref Sagiroglu, Duygu Sinanc, 2013, "Big Data: A Review, In International Conference on Collaboration Technology and System.

[3] Matthew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt, 2012, "Big Data Privacy Issues in Public Social Media", In International Conference on Digital Ecosystems Technologies.

[4] Yuri Demchenko, Canh Ngo, Cees de Laat, Peter Membrey, Daniil Gordijenko, 2014, "Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure", in Proceedings of Secure Data Management Workshop.

[5] Ariana, Mohammad Hossein, and Hassan Rashidi. "Translating from Universal Networking Language into Persian Language." International Journal of Engineering Research and Technology. Vol. 4. No. 05, May-2015. ESRSA Publications, 2015.

[6] Min Chen, Shiwen Mao, Yunhao Liu, 2014, "Big Data: A Survey", Mobile Networks and Applications.

[7] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li , 2012, "Big Data Processing in Cloud Computing Environments", International Symposium on Pervasive Systems, Algorithms and Networks.

[8] Sam B. Siewert, 2013, "Big data in the cloud: Data velocity, volume, variety, veracity", IBM Developerworks.

[9] Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu, Jun Shao, 2014, "Toward Efficient and Privacy-Preserving Computing in Big Data Era", Network, IEEE.

[10] Machanavajjhala, Reiter Ashwin Machanavajjhala, Jerome P. Reiter, 2012, "Big privacy: protecting confidentiality in big data", ACM Crossroads.

[11] LAWRENCE D. SLOAN, 2001, "Echelon and the legal restraints on signals intelligence: a need for reevaluation", 50 DUKE L.J.

[12] Chanmin Park, Taehyung Wang, 2013, "Big Data and NSA Surveillance - Survey of Technology and Legal Issues", International Symposium on Multimedia.

[13] O. Tene, J.Polonetsky, 2012, "Privacy in the age of big data: a time for big decisions", Stanford Law Review Online.

[14] LATANYA SWEENEY, 2002, "k-anonymity: a model for protecting privacy", International Journal on Uncertainty,Fuzziness and Knowledge-based Systems.