Privacy Issues in Big Data from Collection to Use

Alaa A. Alwabel(✉)

King Khaled University, Abha, Kingdom of Saudi Arabia
alwbel@kku.edu.sa

Abstract. Big data management and analysis has become an important matter in academic and industry research. In fact, big data can be very powerful and have significant positive impact on most organizations. However, a large portion of big data in service today is personal data. One of the methods currently in use to preserve privacy of personal data that are collected from big data, is to control the process of data collection according to corresponding privacy policies. In this paper we aim to highlight issues need to be resolved in order to achieve a sustainable balance of growth and protection in the use of big data. We have discussed three main issues including: the advanced definition of personal data, the concern of collecting big data: how and why big data collected and which challenges big data faced, the main privacy principles required in preserving privacy of collected big data, and finally we have proposed a case study to analyze the most famous online organization – Google- privacy policy in collecting big data and decide to which extent privacy is preserved.

Keywords: Big data · Knowledge management · Privacy policy

1 Introduction

Due to recent technological development, the amount of data produced by social networking sites, sensor networks, Internet, online games, online shopping, online education, healthcare applications, and many other companies, is significantly increasing day by day. All the big amount of data produced from various sources in several formats with very high speed is described as big data. Big data has become a very active research area for last couple of years [14]. In fact, big data can be very powerful and have significant positive impact on most organizations. Big data has become the power for better decision making, faster time to market; and enhanced customer service [1]. However, a large portion of big data in service today is personal data. The World Economic Forum describes the personal data gained from big data as—the new oil—a valuable resource of the 21st century, and the analytics of this data as—the new engine of economic and social value creation [2, 3]. Big data about individual including demographic information, social interaction, and internet activity are being collected by different service provider. Personal data is collected and used in order to add value to the business of the organization. The personal data of a person when collected with external large data sets leads to the implication of new facts and secret about that person. This could be done by making insight on people lives without

Y. Tian et al. (Eds.): ICBDS 2019, CCIS 1210, pp. 382–391, 2020.
https://doi.org/10.1007/978-981-15-7530-3_29

their knowing while they might not want the data owner to know or any person to know about them [4]. The other hand, the data overflow presents privacy concerns that could lead to a regularity backlash, moving down the data economy and stifling innovation. Fail to protect data privacy is unethical and can cause harm to data subjects and the data provider. Big Data privacy is the most critical issue since the very early stage of data communication and management. To preserve the privacy of Personal data that are collected from big data, the most efficient way is to control the process of data collection according to corresponding privacy policies [5]. In this paper, we aim to highlight issues that need to be resolved in order to achieve a sustainable balance of growth and protection in the use of big data. We discuss few main questions as following: what is Personal data? Why do service provider collect big data? How do they collect big data? What are the challenges in collecting big data? What are main privacy principles in collecting big data? And finally, we will provide a case study of the most famous online organization -Google- to analyze its privacy policy in collecting big data and decide to which extent privacy is ensured.

2 Personal Data

The usefulness of processing big data is mainly unquestioned, however it arises high privacy risks when working on personal data. This is mostly due to two facts about big data. First, the bigger the amount of data the higher the possibility of re-identifying individuals even in datasets which seem like not to have personal connecting information. Second, it is able to extract from safe personal data new information that is much more important and was not expected to be discovered by the affected person [12]. Personal data meaning is evolving with the advanced technologies to include new dimensions. Traditionally, the definition was pre-determined and governed using personally identifiable information whereas the non-personally identifiable information was often uncontrolled [3]. However, the definition of personal data is changing with new personal preferences, new applications, context of uses, and changes in cultural and social norms. According Teresa Scassa [6], it recognized that personal data means a name, an address and other information that you might give to someone, but now personal data can be any information about our activities, about everything we do online, and even elsewhere [7, 8]. Personal data is a term that may be used in a slightly different manner by different people, but in this paper, we mean by personal data the following two dimensions [9].

2.1 Personally Identifiable Information

Any information that could be used to identify or locate an individual including individual name or characteristics that may be made part of the individual personal data such as age, gender, and relationship status, and it may refer to information that can be correlated with other information to identify an individual like credit card number, and postal code.

2.2 Non-personally Identifiable Information

Any information that could be considered as private and related to the individual life. It could be divided in three sub-dimensions as following.

Social Life. Any information could be used to identify the social life of the individual and considered sensitive such as religion race, health, union membership, personal financial information and job performance information, his interests in hobbies, his interests in entertainment as well as his interests in commercial products.

Cognitive/Expert Life. Any information could be used to identify the expert of the individual including his knowledge, his background and his skills, his professional interests. In addition, individual goal or intention could be also considered to be sensitive and that represent what individual wishes to achieve in a given context.

Digital Life. Any information could be used to identify the individual usage of the digital data, and his unique device identity. The repetitive behaviors of the individual that can be observed and stored in their profiles such as viewing habits for digital content, users' recently visited websites or product usage history, and the history of the individual actions considered under this type of data. Moreover, uniquely traceable to the individual device considered sensitive and may affect individual such as IP addresses, Radio Frequency Identity (RFID) tags, and unique hardware identities.

3 Big Data Collection Concern

Day by day individuals are transacting more and more data online and service provider have begun exploring what insights and lessons they can gather from users of data through collecting, storing, processing, and analysis of extremely large data sets [11]. Although, the collect of data creates influence with each additional use, it leads to the challenges of the complexity in controlling the way of collecting data. Different concerns regarding big data collection including how and why big data is collected, and which challenges does it face will be discussed in this part.

3.1 How Big Data Is Collected

Big Data is collected by billions of connected devices, people and sensors that trace trillions of transactions and behaviors each day. The unexpected amount of data being generated is created in multiple ways. Big Data is actively collected from individuals who provide it in traditional ways by filling out forms, surveys, registrations and so on. They are also passively collected as a by-product of other activities by web browsing, location information from phones and credit card purchases, social media, search engine, free applications and by many other services. The increasing use of machine-to-machine transactions, which do not involve human interaction, is generating more and more data about individuals. All this data is further analyzed and commingled to create inferred data [4].

3.2 Why Big Data Is Collected

According to a report of the standing committee on access to information and ethics (2012) personal data was collected to be used in transaction purpose whereas personal data is now itself the valued commodity [8]. In Professor Scassas view, the data is collected to profile us to define our consumption habits, to decide our fitness for products or services, or to apply price unfairness in the delivery of wares or services. We become data subjects in the fullest sense of the word [8]. Consequently, free online services are not free, but rather a means to commercialize access to users and their personal data. As noted by Jason Zushman of the Merchant Law Group, —the archiving and monitoring of information that's provided by users is what provides the monetary benefits to the companies [10].

3.3 Challenges in Collecting Big Data

World Economic Forum found that the Data-driven opportunities are not without risk and uncertainty. The issue is how to gain new insights and make better decisions, and to do so in a manner that recognizes and protects individual privacy. The profitable incentive in collecting data to share with secondary and tertiary parties is strong and deeply embedded in existing Internet business models. Although, more and more data are collected and combined, the insights, discoveries, value and possible risks increase; particularly, if this activity performed by parties not directly known by individual. With more than 6 billion people connected to mobile devices, more diversity of data is also becoming capable of being linked to individual identity. Smart phones are now able to capture and track an individual location patterns as well as facilitate create new levels of authentication. Moreover, individuals are no longer just the subjects of data they are also being recognized as producers of data. For example, digital personal-health devices measure daily physical activities. They present a new way of capturing a rich data set about an individual [4]. Collecting big data have raised questions regarding the appropriate balance of control between individuals and service provider, and how best to protect personal privacy interests. Researchers argue that individuals have a legitimate interest in the collection of data by third parties. Certainly, big data collection practices rather than bad uses or outcomes are enough to trigger an individual privacy interests. Nowadays, big data collection practices are for the most part unregulated [11]. The key challenge is to regulate the optimal balance between enhanced privacy protection and the helpfulness of the data for decision making. In one hand, the data must be used for extracting value, and, on the other hand, the re-identification of the data must be minimized [13]. Consequently, there is a great need to address the truly activation privacy principles required in collecting big data in order to balance between beneficial uses of big data and the protection of individual privacy.

4 Privacy Principles of Big Data Collection

Principles have been and need to be a core part of the future governance in collecting personal data. Principles can set the foundation for trustworthy collecting data and help empower users. Identifying the principles that reflect communal and cultural norms and ensuring ways to activate them will enable trustworthy data practices, persuading individuals to be more willing to share data about themselves. Existing principles associated with the collection, handling and use of personal data have formed the basis of most privacy and data-protection legislation around the world. Privacy policies must be made available to clients, and be understandable [4]. We briefly describe and discuss the nine privacy principles that should be considered in big data collection process [9, 12].

Notice, Openness and Transparency. This principle refers to that anyone who wants to collect user's information must inform them what and why they want to collect, how they want to use it, how long they will keep it, with whom they will share it, and any other uses they intend for the information. They must also notify users if they want to make any change in how the information is being used. If they want to pass the information to third parties, users must be notified too.

Choice, Consent and Control. This principle refers to that users must be given the choice of whether they want this information to be collected or not. Data subjects must give their approval to the collection, use and disclosure of their Personal identifiable information.

Scope/Minimization. This principle refers to that only information that is necessary to fulfill the stated purpose should be collected or shared. The collection of data should be minimized.

Access and Accuracy. This principle refers to that users must be able to get access to their personal data; to observe what is being collected about them, and to check its accuracy.

Security Safeguards. This principle refers to that safeguards must prevent unauthorized access, disclosure, copying, use or modification of Personal identifiable information.

Challenging/Compliance. This principle refers to that clients must be able to challenge an agency's privacy process. Transactions must be compliant to privacy legislation. One feature of this is respecting cross border transfer obligations.

Purpose. This principle refers to that data usage has to be limited to the purpose for which it was collected. There must be a clearly specified purpose for the collection and sharing of personal data. Data subjects should be informed why their data is being collected and shared at or before the time of collection.

Limiting Use-Disclosure and Retention. This principle refers to that data can only be used or disclosed for the purpose for which it was collected and should only be divulged to those parties authorized to receive it. Personal data should be aggregated or anonymized wherever possible to limit the possible for compute matching of records. Personal data should only be kept as long as is necessary.

Accountability. This principle refers to that an organization must appoint someone to ensure that privacy policies and practices are followed. Audit functions must be present to monitor all data accesses and modifications.

5 Case Study

In our study we are going to analyze the privacy policy of the most famous online organization – Google Company- to address the non-resolved areas in their privacy policy of collecting bid data through online tool that named as policy score. Privacy Score is a project of Privacy Choice, which was founded in 2009 by Jim Brock to make privacy easier for websites, apps and their users. Privacy Score analyses the privacy policies of companies along for clear criteria and assess the privacy risk of using a website. Privacy Score covers two kinds of data: it estimate privacy risk to personal data such as your name or email address based on the published policies of the website, and estimate privacy risk to anonymous data such as your interests and preferences based on the privacy qualifications of the other companies who collect this kind of data across websites [12]. However, Privacy Scores reflects nine factors based on the site's privacy policy and the privacy qualifications of the other companies collecting data there. Four site-policy factors cover how websites promise to collect and handle your personal data. Five tracking data factors cover the privacy policies and oversight of companies that collect anonymous profile data on the site and elsewhere for things like ad selection [12]. Based on these factors, a Privacy Score of 100 would indicate two points as mentioned in the following (Table 1).

Table 1. Privacy score factors.

Privacy Indicator			
How websites promise to handle user personal data		The privacy policies qualification site tracker that collect anonymous profile data	
Factor	Score	Factor	Score
Sharing	30	Anonymity	20
Deletion	10	Boundaries	5
Notice	5	Choice	5
Vendors	5	Retention	10
		Oversight	10
Total = 100			

The Privacy Scores first indicate the site's policies expressly limit the sharing and use of personally identifiable data in four different ways. First, personal data like name, phone number and email address should not be provided to marketers without permission and should be deleted on request. Second, a user's request to delete personal data should be honored. Third, notice should be provided in the case of disclosure of personal data pursuant to legal process or government requests, where legally allowed. Fourth, if service providers have access to personal data, their use of it should be restricted by contract. The other hand, the Privacy Scores indicate that all trackers seen on the site pledge to respect different five criteria. First, personal data should not be collected or use, or should be separated from behavioral data. Second, boundaries should be recognized in areas like health conditions and financial data. Third, choice should provide as to whether data will be collected or applied for the purpose of ad targeting. Fourth, Retention should provide whether data will be deleted data within one year or lose points ratably. Finally, Accountability should be provided through both regular compliance reviews of internal processes by industry organizations such as the Network Advertising Initiative or independent auditors, as well as ongoing external monitoring of practices by industry organizations.

5.1 Google Privacy Issues

Google has drawn considerable criticism for its privacy practices. It were reported that Google used user name and profile pictures in advertisement. In addition, it provides special government access without notifying their user as reported that they participate in PRISM program. In October 2011, it was reported by the United States Federal Trade Commission (FTC) that Google Company are sharing user information without their consent. In August 2012, it was fined $22.5 million by the United States Federal Trade Commission (FTC) for bypassing the privacy settings in Apple's Safari browser in order to track the browser's users and show them advertisements, thereby violating a prior agreement with the FTC.

5.2 Assurances Scale in Google Privacy Policy

Google earns 30 points out of 30 in sharing privacy policy based on the latest updated in their company privacy policy they don't share personal data, like your name, email or phone. Furthermore, they gain 10 out of 10 in their deletion privacy policy as they will delete data promptly when you terminate your account. While they gain zero out of 5 in their notice privacy policy as they don't notify users of government requests for personal data. And finally, they gain 5 out of 5 in their vendor's privacy policy as they require confidentiality from service providers with data access.

5.3 Privacy Qualifications Scale of Trackers on Google Site

Google gain 20 out of 20 in their anonymity privacy policy as they don't associate personal identification data with your profile. Moreover, they got 5 out of 5 in their Boundaries privacy policy as they keep out of sensitive areas like health history, financial records or religion. In addition, Google gained 5 out of 5 in their Choice privacy policy as they allow users to opt-out of behavioral ad targeting. Furthermore, they got 10 out of 10 in their Retention privacy policy as they delete data within one year or lose points ratably. And finally, they gained 10 out of 10 in their Oversight privacy policy as they are subject to oversight from industry organizations. We can summarize Google privacy policy level as following (Table 2).

Table 2. Google privacy policy assessment.

Privacy Indicator			
How websites promise to handle user personal data		The privacy policies qualification site tracker that collect anonymous profile data	
Factor	Score	Factor	Score
Sharing	30	Anonymity	20
Deletion	10	Boundaries	5
Notice	0	Choice	5
Vendors	5	Retention	10
		Oversight	10
Total = 95			

As a result, we can say that during the last two years Google has drawn considerable criticisms for its privacy practices. Although, Google privacy policy has updated to reach better lever and gained 95 scores in both assurance scale and privacy qualification scale for all trackers, Privacy Score has considered other privacy issues of Google Company that could badly affect their users. First, Users will have one profile across all Google services. This means user personal data, searches and behavior are available to all Google applications and advertising. User should log out and use private browsing for Google services that they do not want associated with their Google profile. Second, Concentration of data increases privacy risk. Google's privacy score does not reflect the additional risk posed by the breadth of Google's data collection. If users make extensive Google services–such as Google Search and Gmail - this increases the impact on their privacy if Google does not honor its privacy policies. Third, new privacy policy. This Privacy Score is based on Google's revised and unified privacy policy, which has effective from March 1, 2012.

6 Conclusion

This paper addressed the new perspectives of Personal Data that gained from Big Data and should be protected by service provider. In order to preserve privacy of Big Data, the most efficient way is to control the process of collecting data. Consequently, the paper introduced the way of collecting Big Data, reasons of collecting Big Data and the challenges facing Big Data. The paper also introduced the international privacy principles of Big Data Collection. In order to achieve a sustainable balance of growth and protection in the use of big data, the paper highlighted the no resolved areas in the most famous online organization–Google- privacy policy through analysis tool. Finally, we conclude that privacy principles are not enough and there is a need to translate principles into practice through a model that inform service provider how to manage the use of the personal data in trusted way with economic growth.

References

1. ISACA: Privacy and Big Data (2013). http://www.isaca.org
2. World Economic Forum: Personal Data: The Emergence of a New Asset Class (2011). www.weforum.org/reports/personal-dataemergence-new-asset-class
3. World Economic Forum: Unlocking the Value of Personal Data: From Collection to Usage (2013). www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf
4. Katal, A., Wazid, M.: Big data: issues, challenges, tools and good practices. In: IEEE 6th International Conference on Contemporary Computing (IC3). Department of CSE, Graphic Era University, Dehradun, India (2013)
5. Ng, W.S., Wu, H., Wu, W., Xiang, S., Tan, K.L.: Privacy preservation in streaming data collection. In: IEEE 18th International Conference on Parallel and Distributed Systems. Institute for Infocomm Research, A*STAR, Singapore (2012)
6. Scassa, T.: Standing Committee on Access to Information, Privacy and Ethics. 1st Session, 41st Parliament, May 31, 2012. House of Common, Canada (2012)
7. Pierre-Luc, D.: Privacy and Social Media in the Age of Big Data. House of Commons, Canada (2012)
8. Hasan, O., Habegger, B., Brunie, L., Bennani, N., Damiani, E.: A discussion of privacy challenges in user profiling with big data techniques: the EEXCESS use case. In: 2013 IEEE International Congress on Big Data, Santa Clara, CA (2013)
9. Jason, Z.: Standing Committee on Access to Information, Privacy and Ethics. 1st Session, 41st Parliament, October 16, 2012, 1645 Merchant Law Group (2012)
10. Justin, B.: Why Collection Matters: Surveillance as a Defector Privacy (2012). http://www.futureofprivacy.org/wpcontent/uploads/Brookman-Why-CollectionMatters.pdf
11. Siani, P.: Taking Account of Privacy When Designing Cloud Computing Services. Hewlett-Packard Development Company, L.P. (2009)
12. Jim, B.: Privacy Score Project (2009). http://privacyscore.com
13. Gruschka, N., Mavroeidis, V., Vishi, K., Jensen, M.: Privacy issues and data protection in big data: a case study analysis under GDPR. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 5027–5033. IEEE (2018)

14. Bachlechner, D., La Fors, K., Sears, A.M.: The role of privacy-preserving technologies in the age of big data. In: WISP 2018 Proceedings, vol. 28 (2018). https://aisel.aisnet.org/wisp2018/28
15. Abid, M., Iynkaran, N., Yong, X.: Protection of big data privacy. IEEE Access **4**, 1821–1834 (2016). https://doi.org/10.1109/ACCESS.2016.2558446