# Research Note

## Privacy in the Age of Big Data

**Zhuo Shan**

*Published: May 2018*
*Written:*

Keywords: Privacy, big data, enterprise risk, personally identifiable information, cybersecurity

*Abstract: This paper discusses the various elements surrounding the topic of privacy, particularly in relation to the ever-expanding field of "big data. Part one presents a high-level examination of various techniques used to collect users' data, including an assessment of the definition of personally identifiable information and how corporations obtain it. Part two examines the risks of big data to both enterprises and individuals; part three looks at the applicable government regulations and laws; and the final part provides some risk mitigation recommendations.*

### Introduction

The age of big data is upon us. With the increasing amount of people, devices and services that are connected to the Internet, new data are generated every second from social media, email, search queries and online transactions. As the data mining and analytics technologies advance from the increase in computing power and storage capacity, data have become a new source of economic and social value. [1] Big data technology creates a new raw material for production that drives innovation, efficiency and growth. On the other hand, the extraordinary amount of data collected from users creates concern about privacy that can stir an ethical and legal backlash against the data industry. And the dilemma between big data and privacy will be an ongoing issue for both companies and individuals. We must work together to reconcile the tremendous social and economic benefits of big data with the risks exposed to the privacy of individuals.[2]

### I. PII and Data Collection Techniques

Personally, Identifiable Information (PII) is any data that could potentially identify a specific individual.[3] With the rise of personalization and targeted services, linking data to personal identities has become a key component in many business models, such as online advertising and cloud computing. Therefore, PII has become an immensely valuable asset for corporations. Companies have developed and adopted various techniques to associate anonymous data with specific individuals, which fundamentally undermines Internet anonymization and individual's online privacy. Some commonly used techniques will be examined below.

One of the most common ways to collect user's identifiable personal information is to let them provide themselves. And self-uploaded information can be easily obtained by social media companies and online services providers. User profiles and posts contain personal information, such as birthday, gender, education, family members, etc. By looking at user's content, information about demographics, interests, behaviors and locations can be drawn. [4] Then, companies can create a comprehensive data profile of a user with specific personal details, which can enhance the accuracy of targeted advertisements. Search Engine Companies, such as Google can obtain information from user's search histories; email inbox and cloud drive to target users with personalized advertisements. [5]

Another commonly used technique to track users are Web Cookies. Cookies are small text files that websites place on a user's computer to store certain information about how they use the site. The abilities of Cookies include logging in to sites and delivering personalized content.[6] Cookies not only save login information on a website, they

can also track the users across multiple sites and obtain information. Information is linked to users' identities by identifying their accounts. One of the most commonly used cookies across advertising platforms is Facebook Pixel. It has the ability to track users' behavior when they visit their website again, which would help measure the effectiveness of advertisement. [7]

Most people are aware of the existence of the Web Cookie and its ability. People can simply disable Cookies in their browsers. However, privacy is not guaranteed. Websites can use a technique known as Canvas fingerprinting to identify the users. And it is nearly impossible to prevent it because every device behaves in a subtly different way when websites interact with it. Therefore, third parties can tell when the same device is visiting their websites. [8] Many websites, such as Whitehouse.gov, use this technique to identify users who disable Cookies.[9]

There are many other unique and rarely discussed ways to track users who are browsing websites. One of them I would like to mention is Behavior Analysis. It detects user's physical behavior, such as how user types or clicks, to identify user.

Besides all the techniques used by websites to identify users, there is a much easier way to collect PII -- hiring a data broker. A professional data broker will collect various kinds of personal information of identifiable individuals and sell them to whoever is buying. Companies are buying those profiles because there is certain information they cannot easily obtain online. And data brokers will collect public information such as newspaper subscriptions, financial profile. They can gain insights about your personal life. Once your information is collected, data broker will sell them to buyers, such as your insurance companies, employers or your bank. [10]

## II. Risks and Concerns

Big data poses big risks to both enterprises and individuals. The collecting and analyzing huge amount of personal data implicate growing concerns about privacy invasion, discrimination, information security and government surveillance. This part will examine the risks presented by big data from both enterprise and individual standpoint.

### A. Enterprise Risks

There are serious and potentially catastrophic operational risks for enterprises. First and foremost, information security risk should be treated with great caution. In the past years, there were numerous data breaches happened to companies with sensitive information. In 2013, retail giant Target was hacked and more than 41 million of company's customer payment records were stolen. [11] In 2013, 3 billion Yahoo users' accounts were hacked. [12] Companies with huge amount of personal data have become the biggest targets for cyber criminals. Because of the information security risk, there is also potential legal/liability risk. As we know, not a single firewall or security measure could completely eliminate the possibility of a breach. After the recent cyberbreach, Equifax was hit with at least 23 class-action lawsuits.[13] Therefore, companies have to take in the measure of potential lawsuits after the breach.

In addition, there is also risk of inappropriate business practice, such as invasion of privacy. Because all the private information is completely accessible to service providers due to privacy policy, there is little to no guarantee that companies would be held accountable when they use users' data inappropriately. As a result, companies are more inclined to utilize data with little caution. As reported, Uber CEO Travis Kalanick let his party attendees see all of the Ubers in a city using Uber's "God View" as a party trick.[14] Also, there is the potential of illegal government surveillance. After Snowden incidence, Snowden leaked that NSA was illegally obtaining users' data from companies, such as Google. [15]

Moreover, the ethical concern of discrimination arises because of the increasing usage of automated decision-making process. In a story from New York Times, it is uncovered that Target assigns a "pregnancy predication score" to customers based on purchase habits.[16] Companies use predictive analysis to automatically split users into

categories in order to simplify decision-making process. However, it is problematic and unethical to group people by sensitive information, especially in an equal-opportunity process, such as hiring.

Moreover, failing to mitigate and manage operational risks mentioned above, companies would most likely suffer from reputational loss. Customers would definitely not provide any personal information to a company they find untrustworthy. Without user's trust, any big data companies would deem to fall.

### B. Individual Risks

The risks of big data to individuals can mostly boil down to the invasion of privacy. And the severity of risk exposed to individuals is closely related to the enterprises' ability of managing risks. If individual's data was stolen due to a security breach, identity theft could have a direct impact on one's daily life. There could also be engineered scam based on individual's personal information.

## III. Government Regulations

Currently, there is no single national law in the U.S. regulating the collection, use and sharing of personal information.[17] There are federal and state laws and regulations that apply to certain types of personal information, such as health information. [18] The FTC and White House both released report on Big Data to provide guidance about Bid Data practices. [19]

There is regulation in other countries regarding the practices of Big Data. One of the most notable regulations is EU's General Data Protection Regulation (GDPR), which will go into effect on May 25, 2018.[20] It strengthens and unifies data protection for all people within the EU. And it also addresses the export of personal data outside the EU. This regulation will greatly protect the privacy of EU citizens and residents. There are several key points of GDPR. It gives individual total control over his/her personal data. Individuals will be able to request and remove any personal information stored in the database from any companies operating in EU. It also allows EU citizens and residents to deny the use of their personal information as easily as possible. GDPR also requires companies to provide notification within 72 hours if a data breach occurs. In addition, the fines for violating the regulation are exceptionally high, which could be as much as 4% of companies' preceding years profits worldwide.[21] This regulation as a whole does ensure the online privacy of EU citizens and residences. And it is a great example for the U.S government.

## IV. Recommendations

In this part of the article, I will present my recommendations regarding the risks discussed above. I will discuss recommendations for both enterprise and individuals.

### A. Recommendations for Enterprise

The first and most necessary action for any enterprise with sensitive data is to enhance information security measures. Companies should go beyond compliance, such as PCI DSS (Payment Card Industry Data Security Standard). Compliance does not equal security.[22] Companies should also have an assumption of breach, which means breaches are bound to happen, they are just a matter of time. [23] It is necessary to implement Data Loss Prevention (DLP) technology to monitor and detect potential data breach. After implementing the best information security measures, companies should develop and adopt a recovery plan and follow it if the breach happens.

In addition to information security, companies need to focus on reputation. In order to achieve and maintain positive reputation, companies should start with reducing inappropriate business practices and adopting a transparent business model. Even though there is no law and regulation regarding the use of personal data in the U.S.[24], companies can use EU's GDPR as a reference. The right to access one's personal information is one of the fundamental principles of privacy, which is underutilized in the U.S.[25] Allowing users to access their data will

empower both users' understanding of their data and their trusts in the companies. Moreover, developing and adopting an ethical code of conduct regarding the use of users' data can convey organization's mission, values, and principles to the public, which will provide a positive image for the company.

### B. Recommendations for Individuals

People have different tolerance for risks and different values regarding privacy. Someone values convenience over privacy, and vice versa. However, how companies utilize your data might changes the minds of many people. Companies will use privacy policy as a tool to dig deeper into users' privacy. [26] And there is no way to prevent companies from using that data once you agree to the privacy policy. All the free services are not actually free. We are sacrificing our privacy to access those services. And the real customers who are paying for the services and buying our privacy are advertisers. Companies, like Google and Facebook, are only using the data collected from us to appeal to their real customers. In other words, if you are not paying for it, you are the product.

In order to protect our online privacy, I suggest using privacy tools when browsing online. Privacy tools, such as TOR and VPN, will hide IP addresses and prevent any identification tracking. [27]Encryption, such as OpenPGP, is also a great way to hide your personal information from being accessed.

In the future, when encountering any new service, it is important to make sure that the organization values your privacy and regards keeping your information safe as a mission.

## V. Conclusion

With incredible growth of data technology and increasing use of data, researchers, businesses and individuals are all trying to find a balance point between convenience and privacy. This article has suggested the direction for changes for both enterprises and individuals in the age of Big Data. If organizations ensure the safety of users' personal data and provide users with the access to their personal information, big data will empower everyone and continue to improve society in many ways. In addition, transparency under an ethical code of conduct will motivate users to participate in achieving a greater future of big data.

---

[1] Kenneth Cukier, *Data, Data Everywhere*, The Economist, Feb. 25, 2010, http://www.economist.com/node/15557443

[2] Omer Tene & Jules Polonetsky, *Privacy in the Age of Big Data: A Time for Big Decisions*, 64 STAN.L.REV.ONLINE 63 (2012).

[3] Cory Warren, *What is Personally Identifiable Information (PII)*, Life Lock, Sep.06, 2017, https://www.lifelock.com/education/what-is-personally-identifiable-information/

[4] Jim Matuga, *How does Facebook advertising work*, The State Journal, Dec.04, 2017 https://www.theet.com/statejournal/how-does-facebook-advertising-work/article_f88118dc-641d-5105-add4-4068ff8aa113.html

[5] Google, *We want you to understand what data we collect and use.*
https://privacy.google.com/your-data.html

[6] Inflection, *How Cookies Track You on the Web*
http://inflection.com/blog/how-cookies-track-you-on-the-web

[7] Casandra Campbell, *Relax, advertising on Facebook Just Got a Lot Easier*, Shopify Blogs, Jan.15, 2016.
https://www.shopify.com/blog/72787269-relax-advertising-on-facebook-just-got-a-lot-easier

[8] Simon Hill, *How much do online advertisers really know about you? We asked an expert*, Digital Trends, Jun.27, 2015.
https://www.digitaltrends.com/computing/how-do-advertisers-track-you-online-we-found-out/

[9] *See* note 8

[10] Brian Naylor, *Firms are Buying, Sharing Your Online Info. What Can You Do About It?* , NPR, Jul.11, 2016
https://www.npr.org/sections/alltechconsidered/2016/07/11/485571291/firms-are-buying-sharing-your-online-info-what-can-you-do-about-it

[11] Kevin McCoy, *Target to pay $18.5M for 2013 data breach that affected 41 million consumers*, USA TODAY, May.23, 2017

https://www.usatoday.com/story/money/2017/05/23/target-pay-185m-2013-data-breach-affected-consumers/102063932/

[12] Alina Selyukh, *Every Yahoo Account That Existed In Mid-2013 Was Likely Hacked*, NPR, Oct.3, 2017. https://www.npr.org/sections/thetwo-way/2017/10/03/555016024/every-yahoo-account-that-existed-in-mid-2013-was-likely-hacked

[13] Kevin McCoy, *Equifax hit with at least 23 class-action lawsuits over massive cyberbreach*, USA TODAY, Sep.11, 2017. https://www.usatoday.com/story/money/2017/09/11/equifax-hit-least-23-class-action-lawsuits-over-massive-cyberbreach/653909001/

[14] Kashmir Hill, *'God View': Uber Allegedly Stalked Users For Party-Goers' Viewing Plesure*, Forbes, Oct.03, 2014. https://www.forbes.com/sites/kashmirhill/2014/10/03/god-view-uber-allegedly-stalked-users-for-party-goers-viewing-pleasure/#4e7e6a431411

[15] BBC, *Snowden leaks: NSA 'hacked Google and Yahoo data links'*, BBC, Oct.13, 2013 http://www.bbc.com/news/av/world-us-canada-24753586/snowden-leaks-nsa-hacked-google-and-yahoo-data-links

[16] Charles Duhigg, *How companies Learns your Secrets, New York Times*, Feb.16, 2012. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all

[17] Fred Greguras, *Legal Issues In Big Data: 2017*, Water Online, Jun.29, 2017. https://www.wateronline.com/doc/legal-issues-in-big-data-0001

[18] *See* note 16.

[19] https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf

[20] https://www.eugdpr.org

[21] *See* note 19.

[22] Stephrn Treglia, *Compliant does not equal protected: our false sense of security*. CSO Online, Oct.22, 2015 https://www.csoonline.com/article/2995924/data-protection/compliant-does-not-equal-protected-our-false-sense-of-security.html

[23] UW CISO Office, *Assumption of Breach*, UW CISO Office. https://ciso.uw.edu/about-us/assumption-of-breach/

[24] *See* note 16

[25] Natasha Singer, *Consumer Data, but Not for Consumers,* New York Times, Jul.21, 2012, http://www.nytimes.com/2012/07/22/business/acxiom-consumer-data-often-unavailable-to-consumers.html

[26] BBC, *Google privacy changes 'in breach of EU law'*, BBC, Mar.08, 2012 http://www.bbc.com/news/technology-17205754

[27] Gizbot Bureau, *5 Ways to surf Internet without leaving a digital trace*. GIZBOT, Nov.23, 2017. https://www.gizbot.com/internet/features/5-ways-to-surf-internet-without-leaving-a-digital-trace-045894.html