International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 06 - Issue 02 || February 2020 || PP. 34-39

Privacy Preservation and Data Legacy

Chris Davison [1], David Hua [1], Brady Sheridan [1], Zach Shimer [1], Brianna Bowles [1]

[1]*(Information Systems and Operations Management, Miller College of Business/Ball State University, USA)*

Abstract: Data legacy considerations are often overlooked by consumers. People fail to consider the privacy ramifications and data legacy when allowing their personal consumer activity to be tracked. Groceries purchased with retailer savings cards is an example. Everything bought is recorded in a database, along with any other personal information such as the credit card and personal identifying information. The implications of this transaction raise concern for data legacy, such as how long can this information be held and even used by companies. Individual purchasing behaviors are often aspects people would not want in the public domain, much less for an indefinite period of time.

Keywords: Data legacy, data retention, privacy, privacy preservation

I. Introduction

Often, data legacy is not a consideration when conducting electronic transactions or commerce. It is doubtful that ships' passengers from centuries ago considered the legacy implications of their names being entered in the ship's manifest. Hundreds of years later, those records are public and are preserved in any number of ancestry-related databases. In this research article, the authors explore data legacy implications of apparently benign transactions such as cost saving cards on purchases as well as social media activities.

This research paper begins with a literature review. Following that, definitions and operationalized definitions are presented. A review of the laws concerning data legacy and privacy is then discussed. Then a review of current practices, important occurrences and practices that impact privacy and data legacy are presented. Finally, a summarization and conclusions are presented.

II. Literature Review

Prior research literature fails to directly mention the term *data legacy* but instead covers areas related to the topic. In reviewing the literature of data legacy,three major topics are discussed: legal aspects, end-user agreements and high availability of systems.

The legal ramifications of data legacy are a topic of much-needed research as the laws surrounding the collection and sharing of personal information are examined. There are many existing and proposed laws concerning data collection, data privacy, and security. In this paper, the authors will explore the legal aspects of data legacy.

End-user agreements are another topic for discussion. Users typically agree to them without completely reading or comprehending the verbiage. The result may be sacrificing privacy to use the product.

The last subject presented in the literature review is systems and high availability. Systems and high availability are concerned with how these data collection systems collect, use, and retain personal data.

2.1 Define Data Legacy

There is no formal definition for data legacy. The research literature on the subject encompasses such aspects as retention, usage, visibility, and persistence. Given the lack of specificity, the authors of this research article present the following operationalized definition of data legacy as: the personal data that is collected about a person throughout their lifetime and how it is used or shared within their lifetime and afterward. Personal data can be collected and used in a variety of ways.

The Office of Minnesota Attorney General, Keith Ellison, lists four common ways that our information is collected. The first is governmental records. The government collects data on individuals in a variety of different ways. Some include real estate transfer documents, property tax records, and records from individuals' licenses that are sponsored by the federal or state governments. Databases collect personal information by combining data from several different locations and combining it to create what is known as "People Search Databases". Personal information can be collected by Internet service providers as well. Many states do not have laws that prevent Internet providers from sharing a person's search history with other third parties. The last and most obvious, way our data is collected is through social media and blogs. Most, if not all, social media websites have a terms-and-services policy for joining their website which includes what they can and cannot do with user information. However, this does not stop other companies from using social media from gathering information on individuals for other uses.

2.2 Legal Ramifications

There are a number of laws that concern the intersection of data retention and privacy. Recently in the European Union, the Directive 2006/24 (Data Retention Directive) was repealed by the European Court of Justice based on privacy concerns. [1]

Originally, Directive 2006/24 was adopted to combat terrorism and criminal activity. However, it was challenged and subsequently dropped based on civil liberty concerns. The Directive was found lacking in sufficient control of authorities' access to data and sufficient definition as to whom an authority is and how much data access said authority is granted. This issue is indicative of a larger issue: where privacy ends in the interest of public safety.

Similarly, a Hague court struck down a Netherlands' data retention law [2] for the same reason (i.e., privacy concerns). The law required that all telephone and Internet use be retained for a period of six months, allowing the data to be available for law enforcement purposes. However, the court found that the law did not justify the negative impact on users' privacy.

"Who is Protecting my personal info" [3] states, "In the US, there is no single, comprehensive federal (national) law regulating the collection and use of personal data." In the United States, the laws that cover the collection of personal data are found within other rules and acts, such as:

- The Federal Trade Commission Act
- The Financial Services Modernization Act
- The Health Insurance Portability and Accountability Act
- The HIPAA Omnibus Rule also revised the Security Breach Notification Rule
- The Fair Credit Reporting Act
- The Controlling the Assault of Non-Solicited Pornography and Marketing Act
- The Electronic Communications Privacy Act.

Bartholomew's [4] research discusses several legal and regulatory statutes concerning data collection and retention: The Federal Privacy Act of 1974, The Electronic Communications Privacy Act of 1986 (ECPA), The Stored Communications Act (Title II of the ECPA), and The Computer Fraud and Abuse Act of 1984 (CFAA). Many of these laws were subsequently amended by the USA Patriot Act of 2001 and 2006 as well as the FISA Amendments Act of 2008. Bartholomew indicates a vast amount of disagreement on the amount of privacy either provided or removed by these laws, but most agree that the laws governing the rapidly evolving technology landscape are slow to keep pace.

2.3 End User Agreements

With the advancements in technology and the ability to collect and store significant amounts of data, it is necessary for individuals to be aware of how data is being collected and where it is going. A popular way personal data is being collected is through the use of store reward cards, also known as "money saving cards." BBC News published "*How do companies use my loyalty card data?*" [5] elaborates on this topic. The article hits on how such a card can provide information useful in collecting data for marketing tasks, provides movement tracking, and creates profiles, and shares information with other retailers. This data can be sold to third parties with little to no notice.

It is very common for consumers of such products and programs to be required to agree on an end-user agreement before they can begin usage. This has led to a vast majority of users blindly agreeing to terms without actually reading the document, which covers what data will be collected, how it will be collected, and how it will be used or shared. In a technology-driven world, data collected from different areas can be valuable to a majority of businesses and organizations.

Obar, J. A. &Oeldorf-Hirsch, A. [6] conducted a study in order to explore how many individuals read the terms of service, if at all. The study found that seventy-four percent of people fail to do so, yet they continue to agree to them. In turn, users are then negligent to the information collected and used from the product or service. Knowing the extent of how information is collected and used is becoming a common investigated topic due to the ability to collect the data and individual awareness of such an act.

As mentioned previously, databases are a primary source for personal data collection and sharing of that information, and reward cards are a common way for it to be collected. Again, due to the lack of consumers reading the terms of services and continuing to agree, they begin using the product and potentially provide personal information to a variety of sources. By using Cross-Database Queries, the database administrators (or third parties) are able to combine multiple databases to collect personal information, sometimes for a fee and other times for free.

2.4 Systems and High Availability

Previously, it was discussed how and where personal data is collected. Next, it is important to look into how long the data is kept, as well as the public protection aspect of the laws and regulations surrounding the retention of the data. The laws and regulations about how personal data can be collected and shared covered the public protection side of the law. However, there is another side of laws that cover how long organizations are allowed to hold this data.

According to HIPAA Journal [7], medical health information varies according to state, and can be held anywhere from three to ten years. Credit card companies follow PCI-DSS for their data retention requirements, which states that the retention of cardholder data should be no longer than what is required by law, regulation, or business requirements and after that point should be properly deleted. Medical and credit card industries have well-maintained data retention standards in order to protect the personal data of their consumers.

With that being said, there are others on the opposite end of the spectrum that do not obtain stringent data retention policies. For example, the reward cards previously mentioned have very minimum guidance or standards pertaining to how the time frame their data can be maintained, stored, or used after it is considered no longer useful to the company.

The topics discussed in the literature review bring up several questions and areas of discussion. The authors of this paper will highlight some of the problems noted through this research, raise questions for future discussion, and discuss future research in the area of data legacy and what it means moving forward. Since data legacy is a newer topic or one that is just being recognized, there exist many opportunities to perform research in this domain.

III. Current Issues: Privacy and Data Legacy

3.1 Data Brokers

A data broker collects, analyzes and disseminates information on consumers and organizations. The data that a data broker disseminates may come from public records, the Web, or any number of sales or other commercial transactions. Users that perform these transactions, feeding the data brokers' information coffers, often have no control over their data's dissemination [8]. In many cases, the data broker has no obligation to purge or remove the data unless the user initiates the request through an opt-out process or there are specific data retention laws, such as in the EU, to which a broker must adhere.

Many regulatory acts have been described whose purposes are to protect the confidentiality of personal information. This, however, does not prevent most organizations from selling anonymized data sets. Many organizations have found that their collected data can be a source of revenue [9, 10, 11, 12]. Data brokerages are companies that gather both public and anonymized protected data from as many sources as possible. Data is scraped from social media sites and other public records available online. These organizations take data obtained through these disparate sources and "connect the dots". In so doing, they are able to build comprehensive profiles of individuals that contain confidential and personally identifiable information. An expose on 60 Minutes [13] revealed that lists containing names and contact information of people based on confidential criteria such as psychiatric diagnoses, cancer, sexual orientation, prescribed medications, and much more were available for purchase.

The justification for data brokers is to facilitate targeted marketing. Companies are able to purchase contact lists of individuals who are most likely to be interested in partaking of the product or service that they offer. There is, however, a significant ethical dilemma despite this intended purpose. There are no controls preventing individuals, groups, or organizations from purchasing a contact list for malicious purposes. There is also an expectation that data brokers will retain data in perpetuity. Data brokers with the largest data sets are going to be able to attract more customers and command a higher premium.

3.2 Data Longevity

Related to systems and data high availability, is the concept of data longevity. In the case of data in environments where the retention is unregulated, the cost of purging data may very well outweigh the cost of retaining the data indefinitely. This makes the data purging process a relic of a time when disk space was a substantial investment. Furthermore, when a company is sued, the organization's data pertinent to the case is often requested as part of the discovery process. If the organization is unable to provide the subpoenaed data, that could be quite costly. This gives rise to the phrase: Disk space is cheaper than lawsuits.

Work by Chiou and Tucker [14], indicates that data longevity does not provide any higher degree of accuracy in search result quality. This has policy implications in that organizations often justify their data retention by virtue of search accuracy. Their research also invalidates that argument by demonstrating that older data may inject inaccuracies in searches. [14]

Data longevity may be viewed as a byproduct of modern systems administration and database administration training and practices. High performance and high availability are hallmarks of most systems administration courses [15]. This places an emphasis on availability over time (i.e., longevity). Rowe, Moses, and Wilkinson [16] likened the work of a systems administrator to that of a healthcare primary care specialist. Dealing with aging and longevity is part of the profession. From this, it is reasonable to conclude that database and systems administrators are trained to provide high performance and highly available data in perpetuity.

Relatedly, data longevity may be viewed as a point of practice. The US Census data becomes immediately available after collection. In the US, the population census occurs every ten years. Some countries such as Canada and Ireland conduct a census every five years.

In the US, the census data is scrubbed of personal identifiers. The summary statistical information is made available. However, seventy-two years after each census, the name, and location of respondents are made available.

3.3 Security

The notion of information privacy when dealing with prolonged data retention is further jeopardized when data security is considered. Hackers are constantly seeking out target organizations that may provide personal information that they can sell or use. Even when organizations are diligent in securing their data, the information systems upon which their data reside are still subject to attack by hackers. This has been exemplified repeatedly as the following incidents will show.

3.3.1 Data Breaches

In 2016, Uber suffered a data breach where data from 25 million customers and drivers was stolen. Uber is a ride-sharing company where drivers can pick up and drop off customers similar to taxi services. When Uber revealed the data-breach in 2017, they said the hackers were targeting data being stored on a third-party, cloud-based service. They also said that the information stolen was not trip location history, credit card numbers, bank account numbers, Social Security numbers or dates of birth. [17]

In late 2017, Equifax, one of the largest consumer credit reporting agencies, announced they were also hacked and it just so happened to be one of the biggest data breaches in history [18]. In March of 2017, attackers began searching the web for any server with vulnerabilities that the US Department of Homeland Security CISA Cyber + Infrastructure announced a few days before. On May 13th, the attackers were able to get into Equifax's network using the company's dispute portal. Once inside, the attackers used an Apache Struts vulnerability to gain access to login credentials to three servers. On those three servers, they obtained access to forty-eight other servers containing customer information, at this point the attackers were able to spend nearly two months collecting personal information before being detected.

Lastly in October 2018, the Center for Medicare and Medicaid Services (CMS), a marketplace system for agents and brokers to aid customers in signing up for healthcare, announced they also fell victim of a data breach. The attackers hacked into the government portal that the agents and brokers use to help the customers. The number of affected people was reportedly 750,000. The information that the attackers obtained included names, date of births, addresses, the last four digits of social security numbers, tax filing status, expected income, and additional information. [19]

IV. Privacy Protection

In this section, the authors will discuss aspects of privacy protection. These features and procedures allow users a degree of control over the use and dissemination of their data. Data collecting organizations provide some methods that allow users control over their data and, in some cases, the organizations are required by law to provide these control mechanisms. The authors will begin this section by discussing informed consent. After that discussion, the opt-in and opt-out policy will be presented.

4.1 Informed Consent

Informed consent is the process of obtaining permission to collect and use an individual's data. Typically, this is granted through a signature process (e.g., doctor's office) or the purposeful selection of a check-box or other control box within a web browser. This is accompanied by an explanation of the data use and retention policies as well as any data dissemination policies to which the organization adheres.

As Ingelfinger [20] points out, informed consent is not necessarily educated consent. End users often do not completely read the numerous pages that constitute some informed consent documents. The issue at hand is whether or not the typical end-user can completely comprehend the technical aspects of several pages of informed consent material. Even if the policy documents are completely read, these documents often contain complex legal and technical discussions.

Duress is another aspect of informed consent failure. Informed consent is to be granted freely and under no duress. However, receiving discounts, free items, or other forms of compensation can be construed as coercion. Bechmann [21] claim that social networking sites have degraded informed consent as the use of these sites increases. This is due to social dynamics as well as chasing incentives. While the idea of informed consent is good, the implementation and practice may be lacking in many aspects.

42 Opt-in and Opt-out

The concept opt-in/opt-out is that data collection entities provide all users the ability to begin (i.e., opt-in) and end (i.e., opt-out) of data collection activities. This control is typically coupled with the informed consent policies discussed above.

Lai and Hui [22] discuss the sometimes at-odds aspects of opt-in/opt-out. In their research, they explain that some data collection entities will provide one or the other: opt-in or opt-out. While the two actions may provide essentially the same operative function of user consent, they mechanisms function quite differently in practice. Opt-in requires informed consent and agreement prior to data collection. Whereas opt-out removes already collected data and prevents it from being used in the future. Beales and Muris [23] suggest this approach does not provide users any meaningful control over their consumer data.

V. Conclusion

In this paper, the authors discussed the importance of data legacy as it relates to privacy and individual privacy protection. The authors provided a literature review that defined and operationalized the concept of data legacy. Then the authors provided a discussion of the legal aspects of data legacy as well as a discussion of end-user agreements and a discussion of data legacy as a result of advances of highly available systems.

Following the literature review, the authors discussed contemporary privacy issues with regard to data and data legacy. Data longevity as a result of practice or excellent systems administration leads to private data being accessible for an indeterminate length of time. Accessible data leads to other privacy and security concerns such as data theft.

There have been a number of data breaches and other security matters that have resulted in privacy loss. a number of these hacks and other security problems were presented. Policies and procedures such as informed consent and opt-in/opt-out were discussed. While not a complete solution to privacy and control of data, these practices do provide the end-user with some limited control of private data.

More work is required in the field of privacy as it relates to data and data longevity. Concepts such as Cavoukin's [24] Privacy by Design should be integrated into data systems. Further work on the efficacy of opt-in/opt-out, end-user agreements, and informed consent is required.

References

[1]. Drewry, L. (2016). Crimes without Culprits: Why the European Union Needs Data Retention, and How It Can Be Balanced with the Right to Privacy. *Wisconsin International Law Journal*, *33*(4), 728–754.

[2]. Court Scraps Netherlands' Data Retention Law. (2015). *Information Management Journal*, *49*(3), 12.

[3]. "Who is protecting my personal info?,"(23 January 2019). *PCTouchup Draft*. [Online]. Available: https://www.pctouchup.com/post/who-is-protecting-my-personal-info. [Accessed: 25-Feb-2020].

[4]. Bartholomew, A. (2016). The Smart Grid in Massachusetts: A Proposal for a Consumer Data Privacy Policy. *Boston College Environmental Affairs Law Review*, *43*(1), 79–110.

[5]. BBC News (21, March 2018), "How do companies use my loyalty card data?," [Online]. Available: https://www.bbc.com/news/technology-43483426. [Accessed: 25-Feb-2020].

[6]. Obar, J. A. and Oeldorf-Hirsch, A. (1 June, 2018). The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services. TPRC 44: The 44th Research Conference on Communication, Information and Internet Policy, 2016. Available at SSRN: https://ssrn.com/abstract=2757465

[7]. HIPAA Journal. (18 April, 2019). "Clarifying the HIPAA Retention Requirements," [Online]. Available: https://www.hipaajournal.com/hipaa-retention-requirements/. [Accessed: 25-Feb-2020].

[8]. Tsesis, A. (2014). The right to erasure: Privacy, data brokers, and the indefinite retention of data. *Wake Forest L. Rev.*, *49*, 433.

[9]. Anthes, G. (2015). Data Brokers Are Watching You. *Communications of the ACM*, *58*(1), 28–30. https://doi.org/10.1145/2686740

[10]. Crain, M. (2018). The limits of transparency: Data brokers and commodification. *New Media & Society*, *20*(1), 88–104. https://doi.org/10.1177/1461444816657096

[11]. Matsakis, L. (2019, February 15). *The WIRED Guide to Your Personal Data (and Who Is Using It)*. Wired. https://www.wired.com/story/wired-guide-personal-data-collection/

[12]. Pasternack, S. M. and A., & Pasternack, S. M. and A. (2019, March 2). *Here are the data brokers quietly buying and selling your personal information*. Fast Company. https://www.fastcompany.com/90310803/here-are-the-data-brokers-quietly-buying-and-selling-your-personal-information

[13]. The Data Brokers: Selling your personal information. (2014, August 24). In *60 Minutes*. https://www.cbsnews.com/news/the-data-brokers-selling-your-personal-information/

[14]. Chiou, L., & Tucker, C. (2017). *Search engines and data retention: Implications for privacy and antitrust* (No. w23815). National Bureau of Economic Research.

[15]. López, Pedro, and Elvira Baydal (2017). "On a course on computer cluster configuration and administration." *Journal of Parallel and Distributed Computing* 105: 127-137.

[16]. Rowe, Dale C., Samuel Moses, and Laura Wilkinson., (2015). "Systems administration at the graduate level: defining the undefined." Proceedings of the 16th Annual Conference on Information Technology Education.

[17]. Larsen, S. (22, November 2017) "Uber hack in 2016 exposed data on 57 million people," *CNNMoney*. [Online]. Available: https://money.cnn.com/2017/11/21/technology/uber-hacked-2016/index.html. [Accessed: 25-Feb-2020].

[18]. Cavaliere, V., & Fung B. (2019) "Equifax exposed 150 million Americans' personal data. Now it will pay up to $700 million," *CNN*, [Online]. Available: https://www.cnn.com/2019/07/22/tech/equifax-hack-ftc/index.html. [Accessed: 25-Feb-2020].

[19]. Markham, C. (2019, November 25). "Medicare data breach impacts about 220,000 beneficiaries," *https://www.nbc29.com*, [Online]. Available: https://www.nbc29.com/2019/11/25/medicare-data-breach-impacts-about-beneficiaries/. [Accessed: 25-Feb-2020].

[20]. Ingelfinger, F. J. (1979). Informed (but uneducated) consent. In *Biomedical ethics and the law* (pp. 265-267). Springer, Boston, MA.

[21]. Bechmann, A. (2014). Non-informed consent cultures: Privacy policies and app contracts on Facebook. *Journal of Media Business Studies*, *11*(1), 21-38.

[22]. Lai, Y. L., & Hui, K. L. (2006, April). Internet opt-in and opt-out: investigating the roles of frames, defaults and privacy concerns. In *Proceedings of the 2006 ACM SIGMIS CPR conference on computer personnel research: Forty four years of computer personnel research: achievements, challenges & the future* (pp. 253-263).

[23]. Beales, H., &Muris, T. J. (2019). Privacy and Consumer Control. *George Mason Law & Economics Research Paper*, (19-27).

[24]. Cavoukian, A. (2011). Privacy by design: origins, meaning, and prospects for assuring privacy and trust in the information era. In *Privacy protection measures and technologies in business organizations: aspects and standards* (pp. 170-208). IGI Global.