The Digital Aggregated Self: A Literature Review

Dr. Lynne Y. Williams
Professor, MSIT, School of Information Science and Technology
Kaplan University
Los Alamos, New Mexico, USA
lwilliams4@kaplan.edu

Dr. Diane M. Neal
Assistant Professor, Faculty of Information and Media Studies
The University of Western Ontario
London, Ontario, Canada
dneal2@uwo.ca

Abstract—As the Internet rapidly establishes itself as a major communications conduit, growing concern exists about personal privacy issues and the related ownership of personal data. Privacy and personal data may be vulnerable to exposure by unauthorized individuals, by commercial entities wishing to profit from the data, and even by the individual to whom the data pertains. Although fragments of data may not present a privacy issue on their own, data mining and other aggregation methods quickly assemble data to create a considerably more sensitive “whole.” This article presents an examination of aggregated personal data ownership, or “the digital aggregated self,” using a literature review and an ethical argument. We propose that while server owners may possess the disaggregated user data stored on their servers, individuals should hold the rights to their set of aggregated data that is stored throughout the entire network of online servers.

I. INTRODUCTION

A. The state of digital personal data

Fewer and fewer people today live a life without Internet contact. The ease of online shopping has made purchasing even the most uncommon items as simple as clicking a mouse button. Internet users can manage their bank accounts and credit cards online. They have access to an unprecedented variety of music, movies, television programs, and multimedia content of all kinds. In the course of shopping, managing accounts, or viewing movies, these millions of users leave a trail of personal data crumbs behind them, stored in bits and pieces on every server with which their transactions have exchanged information.

Given that the majority of these commercial servers are generally well secured, does the user have any reason to be worried? The pieces of information stored on a single server may not be a cause for concern, but significant cause for concern may arise when all those pieces are combined using modern data mining techniques.

In 2007, a pair of researchers at The University of Texas at Austin revealed the vulnerability of private data, even data collected in a supposedly anonymous fashion, [1]. Greengard stated that this research “proved that it was possible to identify individuals among a half-million participants by using public reviews published in the Internet Movie Database (IMDb) to identify movie ratings within Netflix’s data. In fact, eight ratings along with dates were enough to provide 99% accuracy, according to the researchers” (p. 17).

The aggregation technique used by the University of Texas researchers is known as a linkage attack. A linkage attack combines individually non-sensitive data with data from other records until a highly accurate aggregated profile of the user is assembled. Linkage attacks are not a new phenomenon. Barbaro and Zeller [2] reported the consequences of releasing half a million AOL subscribers’ search records. Collating the searches made it possible to accurately re-assemble identities, including their names, addresses, and Social Security Numbers.

Supposedly legitimate data mining presents no less of a threat to an individual’s privacy. However, some attempts have been made to protect an individual’s aggregated digital self from data mining exposure. In the United States, this type of safeguard usually falls within the scope of various governmental security standards such as HIPAA, [3], which protects the privacy of Americans’ personally identifying health information. Unfortunately, these types of standards do not exist in many contexts.

Privacy research also offers an array of possible safeguards. For example, Dwork [4], a principal researcher for Microsoft®, advocated a privacy protection method she referred to as differential privacy. Differential privacy does not focus on attempting to prevent all personally identifying information from inclusion in various online databases, which is the more common approach. Instead, differential privacy seeks to preserve privacy by altering the methods by which data is analyzed.

According to Dwork, achievement of “absolute disclosure prevention” is impossible in the practical sense. This is primarily due to the existence of auxiliary information, which is information residing in different locations. For instance, information residing in a hospital database enjoys the protection of
governmentally mandated security standards. On the other hand, auxiliary information residing in search patterns of medically-focused Web sites has no such protection, and can thus be subject to a linkage attack.

We propose a tangential viewpoint to the concept of safeguarding privacy by asking whether the aggregated profile of the individual is “owned” by that individual. A general legal consensus exists which states that the disaggregated bits of data are owned by the owners of the servers on which the data reside. However, as in the Aristotelian saying “The whole is greater than the sum of its parts,” the aggregated whole may represent a body of personal data that could imply ownership. If this implication has substance, the user would have defensible rights over their own aggregation, if not the disaggregated pieces.

II. THE LITERATURE

B. Looking for definitions of digital privacy and ownership

In the course of an average Internet session, a user may be asked several times by many organizations to provide various personal data for many reasons. Perhaps a user is asked to provide a valid name and address to a Web site in order to make a purchase, and then the user may have to provide an e-mail address to another Web site in order to subscribe to a newsletter. As Acquisti [5] noted, in these types of cases, users expect to receive something in return for sharing this data. Further, users retain control of their data to the point of transmittal. Once the data has been transferred to the merchant’s database, the question of ownership of the data becomes clouded.

Parker [6] distinguished between ownership and possession in this manner:

Possession refers to information being in hand or under the control of the possessor but not necessarily known or available for use by the possessor. A possessor is always a living organism (e.g., a person or an animal) and a computer is not strictly a possessor of information. Only the person or persons in possession of the computer possess the information contained therein. In this case, the possessor could be the owner, service provider, custodian, or user of the computer. A corporate or partnership group of people can possess information; but while a government may possess information, it cannot legally own the information since the owners are the people governed and the information is in the public domain. (¶ 3).

This difference between the concepts of possession and ownership is particularly exacerbated when applied to the scattering of personal data across various Internet entities. Using the example of Internet shopping, the user may provide personal information to dozens of merchants who then store the information in databases that reside on physical machines in their possession. Parker [6] noted that in-hand possession of a physical item, such as information typed on paper, is easily established, but “Possession of information in a device that contains other information possessed by others is not so clear. In such cases, possession is determined by who has which types of control. Computerized information is possessed if it is in the full control of the user” (¶ 10). Despite physical ownership of the computer and the database where the personal data resides, the merchants do not necessarily own the data. Parker stated, “Possession does not necessarily result in ownership except as legally stated (e.g., ‘possession is nine-tenths of the law’). Likewise, ownership does not always result in possession under legally or otherwise agreed-upon specified constraints” (¶ 12).

The application of existing laws based on physical ownership provide a certain amount of precedence for establishing ownership of aggregated personal data, but may not conclusively define an owner’s rights. Wamsley [7] described the shortcomings of existing laws:

A property law approach would fail either because the data is not property or because the sender abandons the property at transmission. Privacy law does not suffice, because its provisions are specific and usually relate to personal privacy, not information that has been disclosed or is readily accessible. Contract law may provide the solution at some point, but it would require many contracts between all the various servers and users (p. 3).

If personal data resides on a computer that is not connected to a network and the data is not transmitted (i.e., it does not travel beyond the physical confines of the computer in which it resides), then existing laws can be applied. It is the element of transmittal that begins to change the nature of legal application. In addition to transmission, the element of disaggregation, defined as the “scattering” of personal data fragments across a variety of servers in a variety of locations, creates further difficulties when applying existing laws to establish ownership.

The literature creates questions concerning the very nature of digital data ownership. Can the originator of data be considered the owner? Is ownership implied when it pertains to data unique to or
in identification of an individual? Does transmission and disaggregation change the definition of ownership? Wamsley [7] argued, “A transmission or disclosure likely requires some relinquishment of rights” (p. 8). This concept is most easily illustrated through the concept of jurisdiction. Wamsley [7] noted that various forms of ownership law might apply to data transmitted solely within the United States, but “so much Internet traffic occurs across international borders that these regulations would likely not apply to enough servers to actually achieve data protection either” (p. 10).

Implicit in the definition of ownership is the concept of privacy, or the owner’s right to control access to data which identifies or is attached to that individual. The most common approach taken by commercial entities to overcoming an individual’s concerns about privacy is the use of a privacy policy. As Neal [8] noted, despite the legal weight of The Federal Trade Commission Act in enforcing privacy policies, the policies themselves do not appear to explicitly protect an individual’s data. “Their clearly-defined policies may not sound encouraging to some privacy advocates. For example, Yahoo! states, ‘When you register with Yahoo! and sign in to our services, you are not anonymous to us.’ Amazon writes, ‘You can choose to not provide certain information, but then you might not be able to take advantage of many of our features’” (p. 5).

C. “Legitimate” Aggregation: Data Mining

Databases containing customer information usually hold a catalog of personally identifying information, such as names, physical addresses, e-mail addresses, and so on. For example, if customers purchase items from an online retailer, they are obviously required to provide their home address in order for the items to be delivered. Not all data collection is necessary, however. For example, many Web sites require customers to register in order to view content on the Web site, even if the content is free of charge.

Regardless of the reasons for collection, this information is collected in databases, and then loaded into applications that allow data mining to occur. Data mining can be defined as “discovery of new information in terms of patterns or rules from vast amounts of data” [9], (p. 959). Practical application of data mining abound, such as answering medically-related research questions [10], collecting debts [11], personalizing online user profiles [12], [13], [14], and making informed business decisions [15]. The ever-increasing ability to find accurate patterns with Web-based mining algorithms is staggering; new methods are developed on a seemingly continuous basis. [16] provided one of many current examples.

As might easily be imagined, user privacy is a major concern within data mining practice and research. Any time that data is collected and stored, a chance of inappropriate handling exists, whether intentional or unintentional. Documented privacy breaches by major Internet companies such as AOL [17] and media reports of stolen credit card numbers from e-commerce sites certainly do not calm the worries of privacy watchdog groups or concerned individuals. Poritz [18] rightfully argued that a Web-based search log file breaches the collective community privacy, even if individuals’ data are protected. The handling of sensitive information is of great concern, such as in the case of medical data mining [10], [19]. Additionally, even if organizations strip the identifying information before the data is processed for mining, as demonstrated in [20], it is possible to reassemble the records in a way that uniquely identifies the users [21].

In response to this issue, various approaches to maintaining user privacy within data mining applications have been developed. These efforts attempt to use technology (typically via database algorithms and mathematical models) to protect users’ identities, while still meeting the information needs of the organizations requesting the mined data. In terms of broad categories, the discipline of “inference control” seeks “to prevent published/exchanged data from being linked with the individual respondents they originated from” [21] (p. 452). As previously described, “differential privacy” is a term defined in [1] which refers to the use of mathematics and database algorithms to solve privacy issues [23].

Perturbation methods recombine data so that only newly unique combinations of records exist in the resulting dataset for mining [21]; Aggarwal and Yu [24], [25] discussed weaknesses of the perturbation approach, such as the need to create a new algorithm for each data mining instance. They proposed a new privacy method that migrates the data containing personally identifying information into an anonymized version of the data. Another approach is to store the separate pieces of composite personal data (in other words, related parts of a relational database) in different nodes of the data warehouse [26]. Narayanan, Madria, and St. Clair [27] suggested a system that may provide either exact or approximate answers to queries, depending on the user’s permissions; this approach protects privacy by disallowing access to confidential information.

D. “Not-so-legitimate” Aggregation: Misperceptions and Linkage Attacks

The majority of American web users have an erroneous perception of their privacy rights as protected by legislation in the United States.
There are very few rules in place governing how the data that is collected on you can be used. Privacy is often taken for granted as an inherent right of every citizen, but in fact the U.S. has one of the weakest privacy protection schemes in the developed world. The European Union, Canada, and Japan all trump us in terms of privacy [28], (p. 130).

This lack of a cohesive legal approach to privacy within the United States allows, de facto, any entity engaging in data mining to aggregate data, and use it in whatever way the entity deems fit. Aside from data collection as practiced by commercial entities, ordinary users are likely unaware of the massive amount and detail of data collected, stored and aggregated by major search portals such as Google. In connection with the collection of data by search engines, The Electronic Frontier Foundation [29] reported that “Unfortunately, information stored with a third-party is given much weaker legal protection than that on your own computer. It can be all too easy for the government or individual litigants to get access to your search history and connect it with your identity” (¶ 2).

Taken in fragments (disaggregated), even a large quantity of an individual’s data may not necessarily reveal anything of a sensitive nature. It is the process of aggregation that allows a third party to “connect the dots” and assemble a comprehensive and potentially highly sensitive profile of the individual. [11] provided a frightening example of how data aggregation can be put to use:

It turns out pretty much anyone can set up a collections operation by buying a package of bad debts for around $40,000, hiring collectors who will work on commission, and applying for the appropriate city and state licenses. Once a company is set up it can buy access to Axiom and Experian and other databases and start hunting down defaulters.

So, here we have an entire industry dedicated to buying, selling and mining your personal data that has been derived from who knows where. Even better, because the large credit reporting companies use a lot of outsourcing for data entry, much of this data has probably been processed in India or Pakistan where, of course, the data security and integrity are guaranteed. (p. 37)

Gibbs [11] pointed out that, with no prohibitions on sending data abroad and with the likes of, say, the Russian mafia being interested in the personal information, the probability of identity theft from these foreign data centers is enormous.

Threats to personal data and the databases in which the data are stored do not necessarily come from outside evil-doers; the scenario of the malicious insider is now well-known. As noted by [30], roughly three categories of insider threat exist: naïve users, ordinary users, and malicious insiders. The naïve user may cause a privacy breach by treating data unsafely through ignorance or absence of organizational policies. Ordinary users may unintentionally create a breach through ordinary usage; misaddressed e-mails are a common example. The malicious user needs little explanation; there is usually a monetary motivation, industrial espionage and even loyalty to a nationality other than that of their organization’s location. Before the advent of the Internet as we know it in 2009, Harper’s [31] longitudinal study revealed that up to three-quarters of all American employees have engaged in some sort of “fraud, vandalism or sabotage.” It is doubtful that the proportion has changed significantly since that time.

The fall in the cost of computing equipment and connectivity has made the Internet ubiquitous. People no longer depend on their workplace to provide access to the Web. This increase in access adds exponentially to the population of naïve users and ordinary users provoking unintended consequences. [32] examined the puzzle of why even smart users make bad decisions regarding their data. Taking guidance from research on human decision making, West noted:

Users do not think they are at risk. First of all, people tend to believe they are less vulnerable to risks than others. Most people believe they are better than average drivers and that they will live beyond average life expectancy. People also believe they are less likely to be harmed by consumer products compared to others. It stands to reason that any computer user has the preset belief that they are at less risk of a computer vulnerability than others. (p. 36)

This misconception of risk amplifies the user’s willingness to share personal data in risky ways. The social networking Web site Facebook has become infamous as a source of embarrassing and potentially harmful personal data. Employers are now searching Facebook as a matter of course before hiring new personnel. Banks routinely use Facebook to confirm the trustworthiness of loan applicants. Are these searchers the villains for targeting personal data to their own ends? Not necessarily; the user controls absolutely what they upload and share with Facebook and its 300 million users to date. The user’s misconception of risk makes the user a poor judge of what types of personal data may turn out to be damaging.
Between groups of professional hackers, commercial entities aggregating data for their own purposes, insiders behaving carelessly or maliciously with data entrusted to the organization, and risk-taking users, personal data – particularly aggregated personal data – is under siege. [33] reports that despite the passage of numerous laws in protection of digital personal data, less than 1 percent of the perpetrators are arrested and fewer yet are prosecuted. Law is typically founded on precedent, but the Internet and the availability of personal data residing on the Internet is without precedent, thus making traditional law-making ineffective. Examining individuals’ rights and ownership of the aggregated self through the lens of information ethics may provide a more holistic framework for guidance.

To determine whether an information-based scenario is ethical or unethical, it must be examined in light of Severson’s [34] four principles of information ethics:

1. Respect for intellectual property
2. Respect for privacy
3. Fair representation
4. Nonmaleficence (do no harm)

Severson provided this four-step approach to determine the ethical landscape of an information scenario. First, get the facts straight. Naturally, the individuals represented in the aggregated data would be primary stakeholders. Scenario stakeholders for the aggregated self scenario include commercial entities that would gain from the data in various ways. Criminal entities wishing to profit from identity theft would certainly be stakeholders.

Determining the moral dilemma is the second step. The literature reveals that the majority of Americans do not consider the collection of personal information intrusive if the collecting entity shows good reason for doing so, and if it does not pass the data along to a third party. [35] supported the latter condition, stating that Americans require affirmation that data will not be used in an inappropriate way before disclosure. An example would be the collection of names and addresses by a credit card company. The company is well within its rights to collect this information for the purpose of contacting and billing customers. The dilemma occurs at that point that the company sells the data to, say, a marketing firm. The individual has given authorization to the credit card company to use the data in a specific way. No authorization has been given to either the credit card company or the marketing firm to use the data for other purposes.

The third step evaluates the dilemma using Severson’s principles of information ethics. When the data passes into the control of a third party that the individual has not authorized (such as for the purposes of mining and aggregation), that individual loses the right to fair representation. Individuals may be completely unaware that they are no longer fairly represented. In cases where search engine companies have deliberately released personal data to the public, not only have individuals lost the right to fair representation, but the companies have also demonstrated a lack of respect for user privacy.

In the case of aggregated personal data, the individual has not given explicit authorization to the aggregating entities, whether criminal or legitimate. Aggregation without explicit authorization from the individual described by the data violates the principle of fair representation as well as respect for privacy.

Testing the solution is the final step; will the conclusion bear up to public scrutiny? If the current literature can stand in for public scrutiny, the answer would be a resounding “yes.” Although stated in varying ways, the literature concludes that individuals have the right to authorize how their data is used by every entity accessing (or aggregating) that data. More fundamentally, the literature also concludes that the respect for privacy must be observed by every entity accessing the individual’s personal data.

III. CONCLUSIONS

Some advocates of data mining may argue that data mining itself bears no moral or ethical responsibility to the individual. Data mining as carried out by a network using software has no free will to exercise good or bad intent. This may be a disingenuous argument. Floridi [36] cogently argued in favor of a difference between moral responsibility and moral accountability. Moral responsibility requires “intentions, consciousness, and other mental attitudes” (p. 30) that are naturally dependent on the thoughts of a living being. Moral accountability is not predicated on responsibility but is nonetheless an ethical agent. While mindless machinery and the software that runs it is not capable of moral responsibility, it may still be morally accountable for its actions.

Many instances of data mining transgress at least two of the principles of information ethics, those of the respect for privacy as well as denying the individual fair representation in the use of their own aggregated data. Illegal data mining in the form of linkage attacks violates the principle of nonmaleficence. The individual’s ownership of aggregated personal data can be questioned legally. Ethically, the individual has the right to authorize access to that digital aggregated self
as well as the right to fair representation and the expectation that no harm will result from the aggregation. The disaggregated digital self may not necessarily be owned by the individual, but much more technological, legal, and governmental effort should be made to provide for direct user control of personal data aggregation.

V. REFERENCES

VI. ADDITIONAL READING


P. V. M. Eecke, Truyens, M. "Recent events in EU Internet law," *Journal of Internet Law*, p. 4, June 2008.


