Hidden in the Receipt

Christina Zachary

College of Computing, Georgia Institute of Technology

Hidden in the Receipt

Storming into a Target near Indianapolis, a father angrily waves a book of coupons that had been sent to his daughter by the retail company. “Are you trying to encourage her to get pregnant?” The mailer was full of coupons for baby clothing, cribs, and other pregnancy-related needs. For his well-raised, abstinent teenage daughter, this was a deplorable suggestion on the part of Target and obviously inapplicable, or so he thought. The father received an apologetic call a few days later by the Target location’s manager, only to greet the manager with an apology himself. As it turned out, his daughter was pregnant and could benefit from those coupons (Duhigg, 2012). How did Target know she was pregnant? After knowing this information, is there more to consider morally in blatantly sending consumers such potentially personal, revealing information?

Increasingly a popular technique, companies like Target are collecting data about their consumers, analyzing it, and using the resultant conclusions to make profitable business decisions. Consider the actions taken on an average Monday morning – walking to the bus, stopping by the coffee shop, and lastly, arriving at work. Within this hour long commute, street cameras capture a person’s image as he walks to the bus stop, the RFID enabled bus-card captures who, when, and where a person traveled, the credit card swiped at the coffee shop sends data about what was purchased, when, and by whom, and finally, the company ID scanned upon entering the office sends data as to when and by whom the door was unlocked (Tene & Polonetsky, 2012). This data collection continues throughout the day in a multitude of forms, including online search queries, GPS data, health tracking apps, social media posts and more. By the end of the day, the human race has created 2.5 quintillion bytes of data (Walker, 2015). To put this amount into perspective, these daily records amassed are roughly equivalent to ten

million Blu-Ray discs (Walker, 2015). Broken down, all of these individual bits of data do not leave behind much information to cause concern. However, when the data is collected and amassed by those aiming to analyze it, the potential exists to provide a comprehensive picture of one's life, detailing habits, health records, expected location at any given moment, and many other aspects of personal life (Lazer, 2009).

All of this information, known as "big data," has a serious impact on society's ability to analyze social phenomena, as is occurring in the growing field of computational social science. Previously, computers had been unable to analyze the sheer volume of data being collected, but with the increased processing power developed in the last decade, powerful analysis techniques have been born (Cioffi-Revilla, 2014). Through simulations and models of social data, intricate details of the public can be discovered (Cioffi-Revilla, 2014). The results of these analyses are just what research firms, as well as profitable businesses, are looking to investigate in order to improve upon current research and consumer experience – but at what cost to society? Privacy is a growing concern of many civilians. In early 2015, a Pew Research Center study found 90% of Americans consider controlling what information is collected about them to be important, and most preferred some sort of time limit on how long their records can be stored (Madden, 2015). Currently, no common, widespread, regulatory documents exist to direct the proper collection and treatment of this type of data (Pietri, 2013). The average person has no control over what or when data is collected and cannot remove their personal information from large data sets. Some attempts at creating regulated protections for individuals have been made, most notably in the European Union with the *General Data Protection Regulation*, which outlined principles to follow with respect to consent, disclosure, transparency, and purpose in data collection, among others (Pietri, 2013). Despite these efforts, many computational social scientists and the prying

eyes of large corporations have few guidelines on how sensitive personal data should be collected and analyzed. Many companies have made attempts at being as transparent as possible to please their customers, including explicit verbiage to indicate full anonymization of data. However, recent discoveries are proving this de-sensitized information to not be fully anonymized, therefore exposing personal data (Ohm, 2010). Should ubiquitous data collection continue in the business sector despite the growing privacy concerns? To answer this question, the potential benefits and harms to both data subjects and data holders must be examined.

Benefits

Data Holders. By examining consumer habits and preferences, data holders (the businesses) have the potential to reduce marketing costs while simultaneously creating a personalized consumer experience. Before firms had the ability to mine these rich datasets, the common marketing approach was to send advertisements in mass quantities to all consumers, incurring a large cost which could negatively and significantly impact profits (Acquisti, 2010). The resultant information from analyzing big data can improve marketing abilities by lowering advertising costs and allowing firms to address only those individuals interested in the product being marketed (Acquisti, 2010). These analytics provide the opportunity to target individual types of consumers, delivering them solutions to exactly what they want or need. This approach is evident in the story of the pregnancy-related coupons at Target. The retail company used trends found in big data to correctly identify future parents and provided suggestions for those consumers. This technique has proven its strength in improving both consumer relations and profit. Between 2002 and 2010, after initiating efforts to examine consumer buying habits through big data, Target increased its sales from $44 billion to $67 billion (Duhigg, 2012). This sales boost can be attributed to a minimization of inventory risk, as Target had a better

understanding of its customers' buying habits, as well as a maximization of returns on marketing investment through targeted offers (Acquisti, 2010).

Data Subjects. This reduced cost benefit transfers to the data subjects in the form of time and money saved. Consumers no longer need to search for offers and coupons that apply to them, as these offers are delivered directly to their door (Acquisti, 2010). After backlash surrounding the pregnancy incident, Target continued to identify future parents, but began mixing these baby-care coupons in with other, unrelated coupons to be more subtle in their suggestions (Duhigg, 2012). Target found that, after mixing these coupons, a large percentage of the baby-care coupons were still used, indicating their predictions were correct, and a clear benefit to consumers existed in this personalization of product offers (Duhigg, 2012). While also investigating consumer trends, Kaiser Permanente, a large healthcare provider, found a startling correlation amongst personal data the company had collected. Between 1999 and 2003, 27,000 cardiac arrest deaths could be linked to a drug, Vioxx, which was subsequently taken off the market (Tene & Polonetsky, 2012). The analytics of big data revealed a treasure trove of health information, indisputably benefiting current data subjects as well as potential future patients. In addition to Kaiser's discovery, Google began making use of search queries to predict and locate outbreaks of the flu, ultimately reducing the disease's impact on society (Tene & Polonetsky, 2012). Although the findings may not directly improve profits for the firms, these discoveries do boost the businesses' images in the eyes of consumers by delivering information to them that is both pertinent and valuable, ultimately benefiting both data subjects and holders alike.

Harms

Data Holders. While these health findings illustrate a valuable and rather unexpected advantage from companies examining consumer data, the businesses themselves are creating

competition amongst each other, which may jeopardize innovation (Acquisti, 2010). Collecting vast amounts of information can be costly and unattainable for startups and small businesses. Without access to customer data, smaller firms may find it more difficult to pioneer new products and offer new services (Acquisti, 2010). However, as the cost has begun to decrease over the past few years, companies are increasingly forced into making a decision; forgo the collection of their consumers' data, potentially losing an advantage in the field, or aggregate this data to stay ahead of the industry, taking on the cost and risk associated. Regardless of the expenses to collect data, huge costs can be incurred by companies if the data is breached, handled without care, or is collected in a way the consumer sees as intrusive (Acquisti, 2010). Legal ramifications from a data breach can be detrimental, as illustrated in one such case involving The Home Depot, which was hacked due to a lack of security at self-checkout stands and thus ordered to pay $19.5 million in settlements (Lord, 2015). Since 2005, more than 2,500 breaches have been made public, exposing at least 816 million individual records (Lord, 2015).

Data Subjects. These data breaches not only affect the firms financially and their persona in the eyes of consumers, but also the consumers themselves, exposing them to the possibility of identity theft. Court Ventures, a company which aggregated public records, was identified by the U.S. Secret Service for its reselling of data to fraudulent companies, leading to potentially thousands of cases of identity theft for unsuspecting victims (Lord, 2015). The Court Ventures case illustrates the potential for theft of data subjects' information to not solely lie in hackers' actions, but also in the companies' own decisions. Consumers may trust a firm, providing their information, only to discover later their data was sold to a third party without their permission. Even if this data does not include sensitive information such as financial matters and Social Security numbers, simply the selling of a phone number can be a nuisance –

so much so that recent articles have surfaced informing the general public of ways to avoid getting unwanted telemarketing calls. Anytime someone gives something as simple as his telephone number to a cashier, he is putting himself at risk (DesMarais, 2013).

These failures on the part of data handlers continue further in the anonymization techniques employed. Despite the removal of names and Social Security numbers from data, scientists have demonstrated their ability to identify specific individuals within these supposedly anonymized sets with surprising ease (Ohm, 2010). Anonymization practices have unknowingly failed those who employ them, exposing consumers to fraud as long as data miners know only a few key facts about one's life (Ohm, 2010). For instance, in the mid-1990s, a government agency in Massachusetts released anonymized records of all state employees' hospital visits for researchers to anaylze, removing names, addresses, Social Security numbers, and other identifying traits (Ohm, 2010). Despite reassurance of anonymization from the Governor, some citizens had their doubts. By using voter rolls from the city the Governor lived in, his birth date, gender, and zip code, a graduate student narrowed down the data and found the Governor's private health records with ease, illustrating the fault in purportedly de-identified records (Ohm, 2010). Solutions exist to create data that is more anonymized, although these methods are not fool-proof (Ohm, 2010). A tradeoff exists between the utility of data and its de-identification. As soon as data is made more comprehensive and useful, privacy vulnerabilities skyrocket as more data can be linked in a malicious effort to narrow down and locate individuals amongst the masses (Ohm, 2010).

Moving Forward

The evidence described above illustrates the overt need for corporations to be more careful with big data, despite the benefits it may provide society and how innocuous the data

may seem at surface level. From an act utilitarianism viewpoint, certain situations can be morally justified. Google Flu Trends and the Kaiser discovery of a link between cardiac arrest deaths and a drug are explicit cases justifying big data. However, these wins for society were somewhat unanticipated. Neither of these companies knew when they began collecting consumers' personal data that they would make such beneficial discoveries. What was certain, however, when they began collecting the data, was the risk involved. Overall, firms take on huge risks when amassing databases full of consumer data, jeopardizing both customer privacy and the companies' own well-being. Even if this data results in improvements to society, these benefits seem small in comparison to the worst-case scenario, such as a data breach. What good is a Target coupon for a baby crib when, through sending data to get that coupon, a consumer may experience identity theft, losing much more than the few dollars saved on a crib? The statistics on breaches, misuse of data, and the lack of anonymization paint a dismal picture on the safety of individual data. Either ubiquitous data collection must cease, or alternatives should be found and utilized, which will ensure the anonymization of information and its safety from abuses of power or breaches.

Possible Solutions. Few pathways exist to definitively and irreversibly anonymize data, thus Ohm (2010) suggests all data, no matter how seemingly indistinguishable, should be treated as personally identifiable and subject to regulations either currently or soon to be put in place. However, with this added pressure and liability of ensuring every bit of data is anonymized, organizations may view this task as impossible and be incentivized to abandon attempts at de-identification (Tene & Polonetsky, 2012). In this hypothetical scenario, the effort to ensure individuals' privacy led to an increase in privacy risks, rather than an alleviation (Tene & Polonetsky, 2012).

As an alternative, explicit consent from consumers can be retrieved. Instead of opting-out, individuals could opt-in to collection of their data. While this model does provide the chance for consumers to decide what happens with their information, a switch to an opt-in model can lead to a drastic decrease in societally beneficial statistics. Tene and Polonetsky (2012) cite browser crash reports as an example of this phenomenon. Very few users opt-in to reporting this extremely beneficial information on the presumption that their peers will, and they will be able to free ride on the goodwill of others. While an opt-in approach is a clear solution to retail businesses gathering data for marketing purposes, the lines are blurred when it comes to more societally impactful data, including the information that led to Google Flu Trends and Kaiser's discovery of a drug's side effects. A blanket opt-in approach may thwart innovation, due to an increased focus on direct consent and minimization of data collection (Tene & Polonetsky, 2012).

Rather than change any characteristics of the actual collection of data, a reevaluation of the exchange of data could be examined. Within the medical community, the Health Insurance Portability and Accountability Act of 1996 has been a definite success. Medical records are treated with extreme sensitivity and are not released or exchanged between individuals until a high level of trust is garnered, regardless of if the information is personally identifiable or not (Ohm, 2010). Perhaps this level of trust is what is needed in the world of big data. Corporations and researchers currently maintain high levels of trust with each other, despite the evidence on the contrary, as firms have sold and misused consumer data. Going forward, this consumer data should be treated as the medical community handles its records, moving through review boards and trusting only those certified to handle the information. This method may be the most morally

sound solution to the problems currently plaguing society, although it comes with a high overhead cost, affecting taxpayers and businesses alike.

With respect to the idea of opting-in, a feasible, effective solution has been proposed by researchers at Massachusetts Institute of Technology and the Technical University of Denmark, called "*sensible-data*" (Pietri, 2013). This platform, once completed, will provide improved management of data by consumers and a better understanding of what is happening with their records. "Among the main goals [of the system are] data treatment transparency, fine-grained control in data collection/distribution, and anonymization" (Pietri, 2013). Rather than an all-or-nothing approach currently in use, consumers can choose exactly what data is provided to those collecting and can dynamically change study permissions, constantly informed on the purpose of their information, while researchers can monitor data streaming in or out in real-time (Pietri, 2013). This solution will give the technologically savvy citizen more control and allow the informed citizen to easily opt-in with a few clicks. While this method does have the same downfalls of an opt-in system (less overall data, higher cost), it solves some urgent needs within the big data and computational social science community in a fashion that, morally, continues to provide the data which analysts desire. No longer would every person's information be readily available without their consent, but this data can still be gathered very easily – through a few consenting clicks by the user.

Big data has the potential to unlock insights into society that are otherwise unattainable. "It helps in understanding causes and effects in human behaviour, it gives insights in their interactions, and can explain the inner nature of relationship" (Pietri, 2013). While hugely beneficial, the privacy concerns can undermine the strength of this potentially revolutionary analyses. Before big data and computational social science disappear due to a massive breach,

regulations, or the people's privacy concerns, a solution must be implemented to both alleviate privacy risks and allow these impactful insights to continue. "*Sensible-Data*" is one such solution. Allow consumers the ability to choose whether or not to provide data in an easy-to-use interface and give them transparency into what for and how this information will be used. Use an infrastructure proven to be safe from hackers. After these steps have been completed, and only after, can the field of computational social science continue safely, minimizing risks for both data subjects and holders alike.

References

Acquisti, A. (2010). The economics of personal data and the economics of privacy.

Cioffi-Revilla, C. (2014). Computation and Social Science. In *Introduction to Computational Social Science* (pp. 23-66). Springer London.

DesMarais, T. (2013). *7 Ways Telemarketers Get Your Cell Phone Number | TIME.com*. *TIME.com*. Retrieved 21 July 2016, from http://techland.time.com/2013/07/02/7-ways-telemarketers-get-your-cell-phone-number/

Duhigg, C. (2012). *How Companies Learn Your Secrets*. *Nytimes.com*. Retrieved 11 July 2016, from http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=2&hp

Lazer, D., Pentland, A. (Sandy), Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., … Van Alstyne, M. (2009). Life in the network: the coming age of computational social science. *Science (New York, N.Y.)*, *323*(5915), 721–723. http://doi.org/10.1126/science.1167742

Lord, N. (2015). *The History of Data Breaches*. *Digital Guardian*. Retrieved 12 July 2016, from https://digitalguardian.com/blog/history-data-breaches

Madden, M. & Rainie, L. (2015). *Americans' Views About Data Collection and Security*. *Pew Research Center: Internet, Science & Tech*. Retrieved 11 July 2016, from http://www.pewinternet.org/2015/05/20/americans-views-about-data-collection-and-security/

Marr, B. (2016). *Forbes Welcome*. *Forbes.com*. Retrieved 11 July 2016, from http://www.forbes.com/sites/bernardmarr/2016/01/13/big-data-60-of-companies-are-making-money-from-it-are-you/#3bd2c9964387

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA law review, 57, 1701.

Pietri, R. (2013). Privacy in computational social science. *U RL http://www. compute. dtu. dk/English. aspx. DTU supervisor: Sune Lehmann Jørgensen, sljo@ dtu. dk, DTU Compute*.

Tene, O., & Polonetsky, J. (2012). Privacy in the age of big data: a time for big decisions. *Stanford Law Review Online*, *64*, 63.

Walker, B. (2015). *Every Day Big Data Statistics - 2.5 Quintillion Bytes of Data Created Daily -. Vcloudnews.com*. Retrieved 18 July 2016, from http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/