Data Ethics—Attaining Personal Privacy on the Web

Lisa Singh

1 Introduction

As digital communications continue to increase, people continue to share more and more data, including personal information. As of fall 2015, there were more than 3.2 billion Internet users (Real Time Statistics Project 2015); of these users, 1.5 billion share information on Facebook, 343 million on Google+, 380 million on LinkedIn, and 316 million on Twitter (Smith 2015). Because much of the data shared are publicly accessible, a large opportunity exists for data mining researchers to develop algorithms and methods to support a wide array of analytic services dependent on understanding human preferences. Examples include recommendation systems and customized search tools. On the flip side, the sharing of these large amounts of personal information, some of which are more sensitive in nature, is concerning in the context of personal privacy.

While segments of the population utilize privacy features offered by social media sites, many Internet users do not. Personal demographic information, as well as ideas and thoughts (tweets or messages) that once would have been shared in a more private setting with groups of friends/acquaintances are now accessible to anyone with a computer. If every person and company used these data in ethically responsible ways, then the sharing of so much personal data would be inconsequential. Unfortunately, this is not the case. Individuals, researchers, and companies are using these data beyond their original intent (Tompsett 2005; Hill 2012; Soper 2012; Kramer et al. 2014). The question we have to ask ourselves is—what should we consider reasonable, ethical uses of personal online data?

This chapter begins by discussing some ethically questionable uses of personal data. It then identifies different technologies that are being developed to improve individual privacy on the Internet. The goal of these technologies is to give users

L. Singh (✉)
Department of Computer Science, Georgetown University, Washington, DC 20057, USA
e-mail: Lisa.Singh@georgetown.edu

J. Collmann and S.A. Matei (eds.), *Ethical Reasoning in Big Data*,
Computational Social Sciences, DOI 10.1007/978-3-319-28422-4_7

more control over their data so that the chance of misuse decreases. Finally, this chapter concludes with a discussion about strategies for improving the current situation and suggests the formalization of the field of data ethics to tackle ethical issues specific to the sharing of personal data online.

2 Drawing the Line—Ethically Questionable Studies

Currently, there is no single federal law that adequately regulates the collection and use of personal data from the Internet (Jolly 2015). While guidelines and best practices exist, privacy laws are inconsistent (and sometimes contradictory) across states and dated at the federal level when it comes to limiting use and sharing of personal, behavioral data. At universities, Institutional Review Boards (IRB) have inconsistent standards related to the use of human subject data from the Internet, i.e. ethical uses of available big data about individuals (SACHRP-HHS 2013). Given the inadequate guidelines related to corporate responsibility of personal data and research that uses human behavioral data from the Internet, it is not surprising that a number of ethically questionable uses of data have arisen. We focus on studies from the field of computer science, not because they are more egregious than social science studies, but because they more readily make use of *big data* in their research. This is not surprising since computer scientists can easily manage analysis of large volumes of data obtained from the web. The majority of this section considers two examples, one in cybersecurity and one in data mining. Both of these studies were approved by the researchers respective IRBs. We leave it to you to decide whether or not they should have been.

2.1 Cybersecurity—Planned Attacks and Malicious Software

Some cybersecurity research involves setting up adversarial attacks to better understand insecurities in software. For example, intrusion detection research uses large volumes of network traffic data to generate signatures of potential attacks. Network traffic data contains IP addresses that are not anonymized. However, they are determined using measurement traffic and are not associated with specific individuals (Paxson 2004). Therefore, these analyzes do not violate the privacy of individuals and are important for identifying and preventing different types of cyberattacks. In our viewpoint, this type of research is a good example of ethical big data, cybersecurity research. Of course, we make the assumption that the research follows the federal privacy laws related to access of traffic on computer networks.

There is also a subset of cybersecurity research focused on understanding the proliferation of malware, spam, email harvesting, etc. through the use of botnets. Botnets are a set of compromised machines that can be remotely controlled by an attacker. Because they can grow to be quite large, they are a danger to the Internet community. The most common uses of botnets include Denial-of-Service attacks, spamming, harvesting email addresses, spreading malware, and keylogging (Bacher et al. 2008).

One study used botnets to better understand the "economics of spam" (Kanich et al. 2008). Specifically, they used existing spamming botnets (by infiltrating an existing botnet infrastructure) to understand who and how many people click on spam. They considered two types of spam campaigns, one that propagated malware and one that marketed online pharmaceuticals. To emulate those campaigns and determine click through rates of these forms of spam, the authors created two websites. While the researchers do not actually spread any malware or collect credit card information for their fake pharmaceutical sites, they trick users into believing that they are going to actual sites specified in the spam, e.g. sites where medication can be purchased. Once users click to checkout on one of the websites, an error message is given—no additional personal information is obtained. Neither during or after the process are users informed that they are participating in a study. This could be viewed as a study that manipulates users without their expressed consent. The authors did have IRB approval for this study on the grounds that the authors were not increasing the amount of spam the users were receiving or increasing harm to the users.

While we focus here on only two examples, Burstein has an excellent discussion of legal and ethical approaches for conducting cybersecurity research (Burstein 2008).

2.2 Personalized Data Mining

Numerous success stories involving the use of big data in conjunction with machine learning and data mining have lead to improvements in healthcare, political campaigning, crime prevention, and customer service to name a few (Siege 2013). Unfortunately, there are also examples of researchers and companies using data they have without considering individual privacy.

Many companies use customer purchasing data to send targeted advertising to their customers. While in principle, using internal customer information in this way does not violate any privacy laws, the targeting itself can be unethical. One well known example is Target's marketing of pregnancy/baby related products (Hill 2012; Duhigg 2012). Target determined that when significant milestones occur in people's lives, they are more open to changing their purchasing habits, i.e. switching stores and/or products. Once they make the change, they tend to be loyal customers. Given this knowledge, one campaign focused on the life changing moment of having a baby. Target was able to combine demographic data with purchasing data for approximately 25 products to identify women who were pregnant. Their analytics were precise enough to predict the baby's approximate due date and then market products based on that inference. While on the surface, learning something private about your customers may seem like good customer mining, in this case, Target chose to market coupons to anyone woman they predicted to be pregnant, including teenagers. In one case, they marketed to a teenager whose parents were unaware that she was pregnant. By most people's standards, this is not ethically appropriate. Even if it was a mistake, we need to make sure that companies consider it their obligation to conduct analyses that pass common sense and basic ethics tests.

The final study we consider in this subsection is one conducted by researchers at Facebook and Cornell University (Kramer et al. 2014). To better understand the effects of reading positive and negative articles, these researchers ran a emotional contagion experiment. During a one week period in 2012, Facebook intentionally changed the news feed of over 650,000 random English-speaking Facebook users. For one group they posted news deemed to be more positive. For the other, the top posts presented were more negative. The researchers then measured whether this adjusting of content had an effect on the emotional status updates of the study users. They found that it did (but the statistical significance was small).

The researchers did not obtain consent from the users in the study. Facebook chose to manipulate people's emotions without their consent. The Cornell IRB indicated that the study was exempt from IRB approval since the faculty and student involved did not directly engage with the user data, but instead only had access to the results. The Facebook Terms of Use and Data Use Policy also do not indicate that these types of psychology experiments may be conducted on users of their site. It is a general consent form that does not have the same depth as an informed content document would.

There has obviously been a fair amount of discussion about the ethics of this study (Gorski 2014; Waldman 2014; Chambers 2014). Companies change what we see on their sites all the time. What is concerning about this study is that they chose to knowingly make a subset of their users less happy without telling them. Neither of these studies rise to the negligence of some of the unethical medical studies we have seen, but it is a preview of the types of studies we may see if we do not develop adequate guidelines for human behavioral studies involving big data.

2.3 Online Tracking of Users

A decade ago, online tracking was conducted using simple "cookies" that recorded when a user visited a website and what they searched for on a website. Now, more advanced tools can not only track browsing behavior, but can link that behavior to personal user data including location, demographic data, and even health data (Valentino-Devries 2010; Olsen 2002).

Even more disturbing is that advertisers are paying companies to track people as they use the Internet to better understand what websites they visit or applications they use (EPIC 2015). While this information could be used for targeting ads, it could also be used to target ads in a biased way, that leverages demographic data to `adjust' prices of the same item for different subgroups. In 2010, the Wall Street Journal conducted a study of tracking technologies and found that the top 50 websites installed an average of 64 pieces of tracking technology on their visitors computers, generally, without any warning. Life insurance companies find policies to advertise that fit a user's demographics; health and drug companies map advertising to health terms users are searching for and health related sites they are visiting; and at least one company with social network data is selling it to companies to understand people's creditworthiness—people who are responsible credit users will 'hang out' with other responsible users (Angwin 2010).

Obviously, this type of data collection can violate different Fair Information Practices (EPIC 2015). Users do not know that companies are doing this tracking, they do not know the specifics of the data that is being collected about them, they do not know how the data will be used or with whom it will be shared, and they do not know how accurate it actually is or what inferences are being made with it. Because of these types of tracking software, users cannot make informed judgments about what to share. It also limits their ability to control their data. A need exists to regulate what can be collected and for how long. A need exists for users to be informed about the data values being stored about them and the data values being infered about them.

3 Technologies Being Developed to Improve Privacy on the Web

Most people in the US have a web presence. Obviously, not using the Internet is the safest option, but an unrealistic one in this technological age. While it is unclear how we can improve the ethics of those using large-scale human behavioral data, there are tools available that can make users more anonymous on the Internet and/or can help them better understand the data that companies have about them or that is publicly available on the web. In this section, we consider the types of technologies that either exist or are being developed to help users improve their privacy or better understand their data.

Computer science researchers investigate ways to protect the privacy of user data. In data mining, they focus on privacy preserving methods that hide identifiable information about the user while still maintaining the utility of the data for statistical analysis and data mining. Some recent methods that give companies ways to share or use personally identifiable data without knowing the identity of the individuals include differential privacy for statistical databases (Dwork 2008), anonymization techniques for relational and network data (Zhou et al. 2008; Samarati and Sweeney 1998; Machanavajjhala et al. 2007; Li et al. 2007), approaches for informing users about their Internet data (Singh et al. 2015; Irani et al. 2009), giving users privacy scores to assess their level of vulnerability (Singh et al. 2015; Luo et al. 2009; Gundecha et al. 2011), and prototypes of user controlled identity management systems (Fang and LeFevre 2010; Lucas and Borisov 2008; Luo et al. 2009). While progress is being made, most of these methods are still academic and have not been integrated into real world systems. We surmise part of the issue is that users acceptance of these practices. They are not outraged enough about these practices, so companies have not made data privacy a priority.

Tools that attempt to give users information about their public profile are being developed. One tool that researchers at Georgetown University are working on is part of the Web Footprint project (Singh et al. 2015). This tool constructs *web footprints* of different users by combining publicly accessible information from various online services such as social media sites, micro-blogging sites, data aggregation sites, and search engines about the users. It essentially emulates an adversary searching for publicly available information about a user and has a goal of informing users about data that can be discovered about them. It also recommends the removal of pieces of data that were instrumental in improving the probability of linking and identifying more data attributes. For example, a person's place of work is usually indicative of his or her home state; similarly, a user's home telephone number can be used to infer his or her city. To ensure that adversaries do not use this software, the software only allows authenticated users to check only their names. This software is in prototype phase, but tools like this will be instrumental in helping users understand their public profiles and make adjustments if they choose to.

Tools and best practices also exist to help users reduce the level of web tracking of companies. Here we highlight a few that have been shown to be effective:

- Do not post private information on social media sites. If you choose to, make sure your privacy settings are not set to allow public access.
- Set your browser to not accept cookies from sites you have not visited or sites you do not want to track you. If you do not want to be tracked by the browser itself, some browsers like Chrome have an option for this (incognito mode).
- Do not respond to spam or click on links to sites you do not know.
- Install an ad blocker. This will improve your computer's performance and will reduce data about your click thru habits.
- Referrer data is information that is collected by the previous site you visited. Install a tool to remove referrer data so that other sites cannot access it.
- Encrypt your email so that it can not be viewed by others.

To find other helpful tips, we refer you to (Schmitz 2013; McCandlish 2002; Neagu 2014). These articles describe different types of attacks and possible ways to deal with them.

4 The Pillars of Data Ethics

A survey by Pew Research in 2014 (Madden 2014) showed that most Americans are concerned about data privacy. Over 90 % of adults surveyed agreed that consumers are no longer in control of how personal data is collected and used by companies. Just under 90 % of adult respondents believe it would be hard to remove inaccurate information about them that is online and 80 % that use social media sites are concerned about access to their personal data by advertisers or other businesses. Yet, even with all this public concern, consumers allow companies to do whatever they want with their behavioral data. The idea of not getting free search, Gmail, or Facebook is considered a greater evil than giving these companies free reign on personal data. Until users change their position and hold companies to higher standards, companies will exploit these data as much as they can.

We are at an interesting time—a time when companies and researchers are using technology to drive the understanding of human behavior at an incredible pace. We need to pause and think about what is happening. We need to take back our personal data rights. We need to enforce ethically appropriate use of personal data. It is not big data or big data technologies that are privacy invaders. It is the way people use big data technologies that is invasive and unethical. Here we propose different strategies for improving the current situation.

Regulation. Users cannot regulate companies. Governments need to step up and develop sound regulation about how much and for how long personal data can be collected. Companies should learn from their customer data, but they should do so in a responsible way. If companies cannot do it themselves, then regulations need to be developed.

Data ethics standards. Data ethics standards related to the use of big data need to be developed. There are no safeguards for consumers right now. Because data ethics are complicated, we need a lot of discussion and debate. It is a ripe area for a new discipline to address the complexities that are arising. Data ethics differs from other forms of ethics and needs fresh eyes assessing the moral implications of sharing different types of data.

Catalog of personal data. Individuals need a way to see the data fields a company maintains about them. One way to do this is to setup a mechanism for user to maintain a catalog of the different personal data companies have access to. Users should also have access to new data that is inferred from the original data the company has about them. This is important because the inferences may be inaccurate and users do not currently have a way of knowing that these inaccuracies exist.

Correct inaccurate data. Not all data, original or inferred, is accurate. Therefore, a straightforward mechanism to correct inaccurate data that companies have is important. We can imagine a registry where users have a list of companies that have different data about them and the registry allows users to request companies remove certain data that is too sensitive and/or update it if it is incorrect.

All of these strategies would improve the current situation and allow users to feel more in control of their personal behavioral data.

5 Concluding Thoughts

Companies have a choice about how they use customer behavioral data. Users also have a choice about what behavioral data they share. Unfortunately, the cultural acceptance of publicly sharing personal information on different social media sites and of allowing companies to collect behavioral data without limitations on what they collect or how long they maintain the data for is troubling. The public has been trained that once the data is collected by a company, the company can use it for purposes beyond the original intended use.

This chapter highlighted a number of cases when companies and/or researchers stepped over the boundary of ethically reasonable uses of the data they had. It also highlighted studies that manipulated individuals online without their expressed consent. Finally, it described some technologies that could improve the level of user privacy on the Internet and recommended strategies to help users gain more control over their data.

Every form a user fills out, every click a user makes on a website, every comment or recommendation a user posts about a product, every decision a user makes online is a new data point that is being used by companies and researchers to better understand and potentially infer human behavior. The time has come to pause and debate online privacy and ethical uses of large-scale human behavioral data. The time has come to develop guidelines and regulations that protect users while still allowing companies and researchers the ability to advance knowledge about human behavior in responsible ways. The time has come to take control of our personal data.

Acknowledgments This work was supported by NSF CNS-1223825 and NSF IIS-1522745. The opinions and findings described in this paper are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Angwin, J. (2010). The web's new gold mine: Your secrets. Retrieved November 01, 2015, from http://www.wsj.com/articles/SB10001424052748703940904575395073512989404.

Bacher, P., Holz, T., Kotter, M., & Wicherski, G. (2008). Know your enemy: Tracking botnets. Retrieved November 01, 2015, from https://www.honeynet.org/papers/bots.

Burstein, A. (2008). Conducting cybersecurity research legally and ethically. In *Usenix workshop on large-scale exploits and emergent threats*.

Chambers, C. (2014). Facebook fiasco was Cornell's study of 'emotional contagion' an ethics breach? Retrieved November 01, 2015, from http://www.theguardian.com/science/head-quarters/2014/jul/01/facebook-cornell-study-emotional-contagion-ethics-breach.

Duhigg, C. (2012). How companies learn your secrets. Retrieved November 01, 2015, from http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

Dwork, C. (2008). Differential privacy: A survey of results. In *Theory and applications of models of computation* (pp. 1–19). Springer.

EPIC. (2015). Online tracking and behavioral profiling. Retrieved November 01, 2015, from http://epic.org/privacy/consumer/online_tracking_and_behavioral.html.

Fang, L., & LeFevre, K. (2010). Privacy wizards for social networking sites. In *ACM world wide web conference (www)*.

Gorski, D. (2014). Did Facebook and PNAS violate human research protections in an unethical experiment? Retrieved November 01, 2015, from https://www.sciencebasedmedicine.org/did-facebook-and-pnas-violate-human-research-protections-in-an-unethical-experiment/.

Gundecha, P., Barbier, G., & Liu, H. (2011). Exploiting vulnerability to secure user privacy on a social networking site. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 511–519). ACM.

Hill, K. (2012). How target figured out a teen girl was pregnant before her father did. Retrieved November 01, 2015, from http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/.

Irani, D., Webb, S., Li, K., & Pu, C. (2009). Large online social footprints—An emerging threat. In *International conference on computational science and engineering*.

Jolly, I. (2015). Data protection in united states: Overview. Retrieved November 01, 2015, from http://us.practicallaw.com/6-502-0467.

Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G. M., Paxson, V., Savage, S. et al. (2008, October). Spamalytics: An empirical analysis of spam marketing conversion. In *Acm conference on computer and communications security* (pp. 3–14). Alexandria, Virginia, USA.

Kramer, A., Guillory, J., & Hancock, J. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Science, 111*(42), 8788–8790.

Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k- anonymity and l-diversity. In *Proceedings of the International Conference on Data Engineering (ICDE).*

Lucas, M.M., & Borisov, N. (2008). FlyByNight: Mitigating the privacy risks of social networking. In *Acm Workshop on Privacy in the Electronic Society (WPES).*

Luo, W., Xie, Q., & Hengartner, U. (2009). FaceCloak: An architecture for user privacy on social networking sites. In *International Conference on Computational Science 13 and Engineering (CSE).*

Machanavajjhala, A., Gehrke, J., & Kifer, D. (2007). l-diversity: Privacy beyond k- anonymity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1).

Madden, M. (2014). Public perceptions of privacy and security in the post-snowden era. Retrieved November 01, 2015, from http://www.pewinternet.org/2014/11/12/public-privacy-perceptions/.

McCandlish, S. (2002). EFF's top 12 ways to protect your online privacy. Retrieved November 01, 2015, from https://www.eff.org/wp/effs-top-12-ways-protect-your-online-privacy/.

Neagu, A. (2014). 11 steps to dramatically improve your online privacy in less than 1 hour. Retrieved November 01, 2015, from https://heimdalsecurity.com/blog/online-privacy-essential-guide/.

Olsen, S. (2002). Nearly undetectable tracking device raises concern. Retrieved November 01, 2015, from http://www.cnet.com/news/nearly-undetectable-tracking-device-raises-concern/.

Paxson, V. (2004). Strategies for sound internet measurement. In *Acm Sigcomm Conference on Internet Measurement*. New York, USA: ACM.

Real Time Statistics Project. (2015). Internet live statistics. Retrieved November 01, 2015, from http://www.internetlivestats.com/.

SACHRP-HHS. (2013). Considerations and recommendations concerning internet research and human subjects research regulations, with revisions. Retrieved November 01, 2015, from http://www.hhs.gov/ohrp/sachrp/mtgings/2013%20March%20Mtg/internet_research.pdf.

Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information k-anonymity and its enforcement through generalization and suppression. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*.

Schmitz, D.T. (2013). 5 ways to improve your privacy online. Retrieved November 01, 2015, from http://www.technewsworld.com/story/78590.html.

Siege, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. John Wiley & Sons.

Singh, L., Yang, H., Sherr, M., Hian-Cheong, A., Tian, K., Zhu, J., Zhang, S. et al. (2015). Public information exposure detection: Helping users understand their web footprints. In *International Conference on Advances in Social Networks Analysis and Mining (asonam).* Paris, France.

Smith, C. (2015). How many people use 950+ of the top social media, apps, and digital services? Retrieved November 01, 2015, from http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/.

Soper, D. (2012, April). Is human mobility tracking a good idea? *Communications of ACM*, *55*(4), 35–37.

Tompsett, B. (2005). Identity theft in an onlineworld. *Computer Law Security Report*, *21*(2).

Valentino-Devries, J. (2010). How to avoid the prying eyes. Retrieved November 01, 2015, from http://www.wsj.com/articles/SB10001424052748703467304575383203092034876.

Waldman, K. (2014). Facebook's unethical experiment. Retrieved November 01, 2015, from http://www.slate.com/articles/health_and_science/science/2014/06/facebook_unethical_experiment_it_made_news_feeds_happier_or_sadder_to_manipulate.html.

Zhou, B., Pei, J., & Luk, W. (2008, December). A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM SIGKDD Explorations Newsletter*, *10*(2).