The Ethics of Privacy Strip (Data) Mining

Ethan Johanson
301076978

Privacy is a term that can be interpreted in a wide variety of ways by different people, and even to the same individuals based on its context. In some circumstances, these definitions will align between a society's legal understanding and the individual's assumptions, but in many cases and situations these two paradigms are not necessarily in agreement. In the modern digital age, these issues have become compounded as means of gathering and recording information have dramatically increased in both availability and lowered the difficulty of being performed. As technologies have evolved, so too have society's understandings of what privacy entails and what an individual can reasonably expect out of various situations. In some cases though, such as data mining, what the average person expects of the information they have chosen to share ends up being far from the end result that information may provide. There are instances where businesses may sell this information gathered about their customers, with its potential for the uncovering of further information, to other third parties. Whether this discrepancy in expectations can rightfully be considered a violation of privacy rights is largely a contextual question based on the parties involved and the agreements set upon during the initial exchange of information. The ethical implications of such a situation is also somewhat contextual, but arguably fall mainly on the side of the behaviour being unethical, excepting extremely unique circumstances.

The ubiquity of portable recording devices with mass storage capabilities combined with the plethora of mediums available to distribute the recorded information has resulted in a situation unseen in any prior point in human history. A simple out of the ordinary action performed in a public place can be filmed by a passerby and uploaded, unbeknownst to the subject of the recording, and ‘go viral’, resulting in online fame (or infamy) and recognition as being the individual in the video by random strangers as well as peers long after the initial incident. This sort of perpetuity of moments is not entirely novel. It is without question that one could gain a reputation within their town through the spread of gossip, albeit without documented evidence and simply by word of mouth. The classic image of a western wanted ad hanging in the saloon is another contextual foundation of an individual’s image being manipulated and controlled by outside forces, but again, these mediums required some level of effort and disseminated at a much slower pace than the world wide web easily facilitates. It can be logically concluded that the spread of less complex information contained in a simpler text-based format is just as easy, if not simpler, to have reach interested entities. A secondary element that one must be aware of when considering the implications for privacy in the digital age aside from the ease of information distribution, is the ease of information analyzation. The practice of data mining has its roots in the field of statistical analysis, which does not require computational power. There is little argument however, that the addition of the computer as a rapid calculation and aggregation tool has dramatically altered the ease, scope, and interconnectedness of such practices. While before a trained professional may have pored over charts, marking locations and events, creating a web of information to derive a hypothesis or conclusion on the details of the information or attempt to predict future data, this process would take time and a careful mind to make sure that nothing was forgotten or misremembered. This sort of visualization and analysis is often represented in media incarnations of the tale of Sherlock Holmes. What was once the realm of what a theoretical mastermind could achieve over months has become possible in a matter of minutes with the advent of machines capable of performing millions of calculations a second. When this processing and analytical power is harnessed in conjunction with the advanced information accessibility and distribution afforded by modern technology, the difficulty in generating a comprehensive package of information on a individual or group of individuals becomes exponentially simpler. It is with this background of understanding and context that one begins to appreciate the scale of the difference in one's control over their public image and available information between now and the majority of history. With such a gap in place, it is understandable that one would be concerned with the ramifications and ethical implications of such a change in activity.

These ramifications and implications will vary based on the issue at particular hand and the positions of stakeholders involved. For general purposes, one can look into statistics provided by research into privacy habits. Rakesh Agrawal and Ramikrishnan Srikant's article, *Privacy-Preserving Data Mining*, references the results of a recent-at the time survey, which classified 17% of respondents as privacy fundamentalists who would not provide data to a web site despite any privacy protection measures being in place, 56% as those whose concerns were significantly reduced by the presence of privacy protection measures, with the remaining 27% marginally concerned and generally willing to provide data to web sites. They also provided another survey which found that 86% of respondents believed that participation in information-for-benefits programs was a matter of individual privacy choice. This included 82% saying that having a privacy policy would matter, with only 14% stating that it was not important as long as they received benefit. It was also found that people were not equally protective of every field in their data records. [1] From these statistics one can see that the option of having privacy and the concept of a predefined agreement upon the definitions and rules of information exchange is a highly popular sentiment. There is also a significant group of individuals who comparatively place little personal value on such ideas, for whatever their reasons may be. Siva Vaidhyanathan, in his article *Naked in the 'Nonopticon'*, provides a postulation on the generational gap in attitude towards privacy seen in the modern Facebook culture, courtesy of journalist Emily Nussbaum. *"Younger people, one could point out, are the only ones for whom it seems to have sunk in that the idea of a truly private life is already an illusion," Nussbaum wrote. "Every street in New York has a surveillance camera," she said. "Each time you swipe your debit card at Duane Reade or use your MetroCard, that transaction is tracked. Your employer owns your e-mails. The NSA owns your phone calls. Your life is being lived in public whether you choose to acknowledge it or not."* [2] This observation draws upon the idea that those who have grown up with current technology have a greater understanding of the realities of that technology's scope, and have in many ways simply resigned themselves to a state of acknowledging privacy as being something that no longer truly exists. Vaidhyanathan captures this thought in his article, pointing out *"despite warnings from nervous academics and almost weekly stories about extensive data leaks from Visa or AOL, we keep searching on Google, buying from Amazon, clicking through user agreements and privacy policies (which rarely, if ever, actually protect privacy), and voting for leaders who gladly empower the government to spy on us."* [2] After noting these elements of the current era, Vaidhyanathan goes on to postulate that our understanding of people's attitudes towards privacy is flawed based on the wide interpretations possible of what exactly privacy consists of and what it means. He suggests that at least in a social media sense, the information itself is less relevant than who retains the power to control, collect, and share that information, referencing the outcry in 2007 when Facebook introduced a social advertising program placing users' purchases in newsfeeds. He goes on to suggest a new method of thinking about privacy, defining it within terms of four distinct domains. These include person-to-peer, person-to-firm, person-to-state, and person-to-public. Of particular note is the analysis of person-to-firm, stating *"We gladly accept when we are offered what we think are "free" services, like discount cards at supermarkets and* *bookstores that actually operate as record-keeping account tokens. The clerk at the store hardly ever explains how the system works or what the nature of the transaction really is. We don't always realize what we are giving away when we thoughtlessly reveal a simple piece of data like a phone number or ZIP code"* [2] This observation highlights the tradeoff of benefit for information decisions that many people make every day, recalling the 14% of individuals from the study referenced by Agrawal and Srikant that did not even care about a privacy policy if there was a benefit for them. Overall, a general trend can be seen that with the lure of enhanced benefits and the idea that most information is already available, people are readily giving up more and more formerly private information to the public domain. In some cases, such as Google's recent privacy policy changes, which integrates profile information across multiple services including Search, GMail, and YouTube to provide a more detailed and comprehensive user profile, there have been mixed reactions, ranging from calls that the policy violates European data protection legislation [3], to a feeling that the outcry is a large overreaction to a change in perception rather than any actual legal changes, and that as been the case the entire time, individuals are perfectly free to opt out of the benefits google provides in exchange for the information it gets from its users. [4] Again though, these cases are largely confined to the realm of particular pre-agreed upon information types, while the practice of data mining opens a whole new world of potential areas of dissent.

In his paper *Informational privacy, data mining, and the Internet* [5]*,* Herman Tavani argues that the practice of data mining differs even from now traditional practices of retrieving information from computer databases. He sums these key differences up in six ways. The first is that implicit patterns can be derived via data mining as compared to the explicit nature of information extracted from a traditional database system. This allows associations to be discovered rather than requiring them to have been provided beforehand. Second is that data can be compiled and utilized from a single database (or data warehouse) rather than exchanging information between multiple databases, thereby bypassing the need for explicit consent to have information available to be shared across databases, and increasing the efficiency of information discovery. The third trait is the utilization of open-ended queries to discover information on relationships and associations vs explicit queries. Again, this facilitates the uncovering of information that otherwise would not have been apparent. Tavani's next point is tightly related to the previous, in that the information from data-mining has a non-predictive aspect to it, where the organization doing the data mining does not know beforehand what sort of use they may gain out of a search for interesting data, as opposed to a more traditional setup, giving the example of a law enforcement agency looking for particular matches to simply check if an individual is listed within two predetermined databases. The fifth point raised is the relative public nature of the information being mined. What this point attempts to highlight is that as opposed to more confidential traditional database information such as health or financial records, many currently-mined datasets involve public behaviours such as shopping habits that were traditionally in a domain where actions had far less expectation of having any permanence. Tavani's final core way that data-mining differs from traditional practices is the ability to construct new groups or categories of persons based on patterns as opposed to extracting information on single individuals. Unlike traditional group identifying methods, the data manipulator needs no preconceived notions of the type of group and parameters they want to look for, and rather can have huge amounts of data parsed to generate entirely new types of groups that may have been formerly overlooked. Tavani goes onto illustrate these elements in practice, examining the hypothetical situation of a bank customer providing some information to their bank, and subsequently being identified along with other clients as belonging to a new type of high risk group that was previously unknown, likely even to the customer himself. With the bank not having specified when collecting information for the loan that the client's information could be used in these alternative practices or obtaining permission to do so, there are questions raised as to the responsibility of the bank to either not perform such data mining or to warn all users beforehand of the various ways their data may be used. With the unpredictable and open-ended nature of data-mining however, it is arguable that the bank could never properly inform clients of what may be done with their information. Taking these differences into consideration, one can see the dramatic gap between the implications of traditional information gathering and the results of deep data-mining techniques. It is also worth remembering, as John Jones writes on his short post *Digital Literacy: Search Algorithms are Mechanical Turks* [6], that there is a"*common assumption technological processes aren't capable of deception or other forms of obfuscation. That is, machines, they don't lie. They have inputs and outputs, but the quality of the output depends on the input, not on the algorithm that processes it.",* but that in actuality, *"algorithms, even Google's algorithm, depend entirely on the biases and decisions made by their creators.... even though algorithms give the appearance of autonomous behavior, maintaining that semblance of autonomous action requires frequent tinkering and is dependent on guidance from human controllers. In this sense, algorithms are mechanical, in that they have features that enhance the speed or accuracy of these human decisions, but they are not independent of those decisions."* While Jones is primarily writing about search algorithms, it is also true of any data-mining algorithm. Despite any and all intentions to create a perfect algorithm, certain choices will have been made by the software engineers and these choices will be reflected in any findings made by the software system.

The sale of customer data to third parties then, is a particularly powerful process that facilitates a large array of possible outcomes based on who is processing the data, what the data is, and what the processors are attempting to find. In many cases this third party is indeed a marketing agency attempting to refine their target markets or create more effective advertising. As noted by Tarvin, these intentions are not necessarily the results that will be created from such acts of data mining, which precludes customers being afforded the right to know the ways in which information they are providing will be utilized. With proper anonymity measures in place though, it could be reasonably argued that the remaining data is purely information the business would understandably expect to know about themselves and how they interact with customers, rather than how customers react with them. From this angle it may be difficult to argue that consumers' privacy is being violated. In the hypothetical case of a restaurant for instance, the company could look at their data of how many types of their various offerings sell at different times and in what combinations, whether certain meals produced better tips, or a variety of other potential factors that are quite far from being information that is tightly coupled to individual customers. It is reasonable to expect a business or individual to desire to and be able to assess their success in dealing with others. Once this sort of information has been passed to a third party for analysis though, it may also be expected that such is the extent of the reach of this information, as any further sharing by the third party has removed both parties of the initial transaction from the equation, resulting in an entirely new set of stakeholders who may have a different sense of what this information can and should be being used for, despite having no claim to its creation.

Based on all of these contexts, one can begin to form a theory on whether the practice of selling information to third parties violates privacy rights of the parties to the transaction.

Recalling Vaidhyanathan's proposals [2] and Agrawal + Srikant's referenced findings [1], one can see that people's expectations of privacy can not be applied in a consistent manner across either contexts or individuals. If one does not possess expectations that a piece of information will be private or to be subject to privacy rights in the first place, is it possible to break that individual's privacy rights at all through use of said information? Depending on the context of how this information will be used, Tavani [5] would likely suggest that such a thing is indeed possible. With advanced modern data mining techniques, information can be discovered that neither party taking part in the original information gathering would expect to exist. As this newly discovered information was not previously known, any and all decisions of the individuals in making the choice to share information in the first place must have been made and decided upon without knowledge that this uncovered element was included in the transaction alongside the originally supplied data. By adding this additional element to the equation, the fundamental ideal of an individual being able to choose how they present information about themselves and with whom, particularly in the peer to firm context, has been averted. When this is combined with the addition of a third party and potentially more to the list of stakeholders in the information, the issue rather quickly reveals itself to be a breach of privacy rights, even if minor at times. However, with a comprehensive privacy policy, a business could quite likely be able to defend itself from these charges, provided the document any customers agree to contains a description of all of the methods of any data mining that may be utilized with the specific customer data that will be being sold and used. With such information available, there is little a consumer could do to argue that rights were violated, as they voluntarily entered into an agreement outlining this exact potential.

While legal rights are one thing, the ethics of the situation are an entirely separate domain that does not require the agreement of the former to arrive at its own conclusions. Depending on one's ethical belief system, there are multiple assessments that could be arrived at. A follower of Kantianism for instance, would almost certainly end up determined that the practice is unethical. This follows the first formulation of Kant's Categorical Imperative, "*Act only according to that maxim by which you can at the same time will that it would become a universal law."* Were one to attempt to universalize the unprompted disclosure of privately entrusted information to others, this would not result in a very functional world and would effectively render the idea of confidentiality a thing of the past. The practice is in contrast even greater to Kant's second formulation, *"act so that you treat humanity, whether in your own person or in that of another, always as an end and never as a means only".* [7] In taking information about customers and using it as a means of generating output without their consultation and approval, be it payment from the sale or extraneous information to utilize, the business would definitely be responsible for utilizing their customers as a means to an end rather than as an end themselves. From contrasting the practice with these formulation rules, it is apparent that from a Kantian view, such sales are unethical. For a utilitarian perspective, one is concerned with attempting to produce the maximum utility, or 'happiness'. Two main branches of this ethical philosophy are Act and Rule Utilitarianism, which are largely divided along calculating utility from a specific decision and calculating utility from a general rule for that type of decision. In this case one may identify the stakeholders as the initial customers, the business the customers dealt with and its shareholders, and the third party who is being sold the information. The action of selling information will almost certainly produce a net 'happiness' increase for the initial business, as they are provided with a new source of income and potentially learn new information to gain more profit with in the future. The third party gains a similar increase in happiness through having their services used and appreciated, and potential insights gained and skills refined through getting to analyze/utilize the sold data. The customers are a more difficult group to assess however. Amongst the consumer group there will almost certainly be a split between those upset with their information being distributed and those who do not care. The latter category though, is not producing any positive happiness gains to counteract the effects on their peers who are losing happiness. There is also the potential to consider that this lost happiness group could sharply increase were there to be an information leak that released the gathered and mined material much more publicly than they expected based on their initial contexts of their interactions. As one can assume that the ratio of customers to business members is fairly skewed towards the consumers being the larger group, it is a likely result that in terms of rule utilitarianism, the sale of information results in a near net zero increase in 'happiness' at best, with potential for a huge downfall in happiness were circumstances to go wrong. Weighing this cost benefit analysis against the neutral lack of information selling shows that as a rule, it is more prudent to optimize utility through not selling information to third parties. Following Act Utilitarianism, one can imagine specific scenarios where if the business has a strong and well supported suspicion that the sale and subsequent analysis of their data could result in a breakthrough that provides a large societal benefit, say in the field of health, that the sale of that information could conceivably be ethically justifiable.

Overall, from looking at the state of technology and its potential in the modern age, and how it has affected the public's perception of privacy, there is a sense that a shift is indeed occurring in the domain of how one controls the private elements of their lives. From examining the potential in data mining in particular, it is notable how predetermined agreements of how information may be used when recorded in interactions between individuals and other entities such as businesses or the state must become more and more comprehensive to cover all potential outcomes and uses of that data in order to assure people that they retain some control and choice over how their interactions are presented to the world and to maintain legal protection from claims to violations of rights. The implications of selling customer data to third parties may indeed be covered in terms of privacy rights thanks to such practices, but from a purely ethical standpoint, both Kantianism and Utilitarianism can be used to argue against the concept, excepting in highly unique and specific cases for one branch of Utilitarianism. Due to this somewhat rare form in agreement between a rule system based on consequences and one based on intent, it is not a stretch to claim that this sort of sale can be reasonably claimed as an unethical practice.

References

[1] Rakesh Agrawal and Ramakrishnan Srikant. “Privacy-preserving data mining”, “*Proceedings of the 2000 ACM SIGMOD international conference on Management of data*(SIGMOD '00).” [Online]. Available: http://doi.acm.org.proxy.lib.sfu.ca/10.1145/342009.335438 [Accessed Mar. 1, 2012].

[2] Siva Vaidhyanathan. “Naked in the ‘Nonopticon’”. “The Chronicle Review Feb 15 2008.*”* [Online]. Available: *http://www.google.ca/url?sa=t&rct=j&q=naked+in+the+nonopticon&source=web&cd=1&ved=0CCUQFjAA&url=http%3A%2F%2Fwww.english.illinois.edu%2F-people-%2Ffaculty%2Fdebaron%2F582%2F582%2520readings%2Fsiva%2520on%2520privacy.pdf&ei=ay1XT8HpI6KOigKowM3LDw&usg=AFQjCNEtkFDl9re_49YT63E63Md4Tawtpw*

[Accessed Mar. 1, 2012]

[3] The Guardian, “Google privacy policy changes spark Europe-wide inquiry” 2012. [Online]. Available: http://www.guardian.co.uk/technology/2012/mar/01/google-privacy-policy-changes-eu [Accessed Mar. 3, 2012].

[4] Jon Mitchell. “Tech World Overreacts to Google’s New Privacy Policy - How Does It Affect You?”. [Online]. Available: http://www.readwriteweb.com/archives/tech_world_overreacts_to_googles_new_privacy_polic.php [Accessed Mar. 3, 2012].

[5] Herman Tavani. “Informational privacy, data mining, and the Internet”. [Online]. Available: http://www.springerlink.com.proxy.lib.sfu.ca/content/v355v06379533776/fulltext.pdf [Accessed Mar. 1, 2012].

[6] John Jones. “Digital Literacy: Search Algorithms are Mechanical Turks”. [Online]. Available: http://dmlcentral.net/blog/john-jones/digital-literacy-search-algorithms-are-mechanical-turks [Accessed Mar. 1, 2012].

[7] James W. Cornman, Keith Lehrer, George Sotiros Pappas. “Philosophical problems and arguments: an introduction”. [Online]. Available: http://books.google.ca/books?id=cRHegYZgyfUC&pg=PA336&dq=categorical+imperative+kant+means+to+an+end&hl=en&sa=X&ei=_IFYT7_FCuiXiALYvPGTCw&ved=0CDAQ6AEwAA#v=onepage&q=categorical%20imperative%20kant%20means%20to%20an%20end&f=false [Accessed Mar. 3, 2012].