Data Discrimination

The Importance of Keeping Confidential Data Safe in the Hazardous Data Mining Work Environment

Edward A. Lockard

West Virginia University

December 13, 2009

"The human brain contains roughly 300 megabytes of information. Not much when you get right down to it. The question isn't how to store it, it's how to access it. You can't download a personality. There's no way to translate the data. But the information being held in our heads is available in other databases. People leave more than footprints as they travel through life: medical scans, DNA profiles, psych evaluations, school records, emails, recording, video, audio, cat scans genetic typing, synaptic records, security cameras, test results, shopping records, talent shows, ball games, traffic tickets, restaurant bills, phone records, music lists, movie tickets, TV shows… even prescriptions for birth control."

- Exert from the pilot episode of *Caprica*

The aforementioned quotes come from the pilot of a science fiction television series to debut in January. As this is a work of fiction, this situation of a digital copy of one's personality being created from a simple join between multitudes of databases is not the focal point of discussion. However, the concept which it explains in more simple terms does present a scenario which would be considered a major affront to privacy advocates: an individual or organization, whether it is in the private or public sector, possessing nearly all information on an individual. In the context of the quote, the data is being used for positive purposes as a way to live on after death. However, this paper will discuss the possibilities of private information being used by a party to cause harm to an individual, be that physically, financially, or even legally.

Now, some may question why privacy is such a hot-button issue in the United States and other democratic country. One could argue that law abiding citizens should have nothing to hide. Another argument could be that if several collections of information are matters of public record, there is no reason that they should be allowed to be made into one large electronic record. Both of these arguments do possess a degree of validity, especially after the events of September 11th. However, for as beneficial as Data Mining concepts could be in fields of research, for example, these same concepts could have an even greater negative impact if applied for the wrong purposes.

One negative consequence that access to too much of our information by the wrong parties is discrimination. Usually when one thinks of discrimination, one thinks of a person being treated unfairly based on race, sex, or nationality. However, when one thinks of those characteristics simple as pieces of information, discrimination could almost be redefined as an unfair bias based on a on a certain set of attributes. Before

discussing possible occurrences of "data discrimination" in a more legal context, let us discuss the idea of using one's data to discriminate against them, financially.

In Joe W. Pitts' article for *The Washington Spectator* titled "The End of Illegal Domestic Spying? Don't Count on It," Pitts makes mention of reasonable innocent application of our data on shopping websites, such as Amazon.com. Using our search logs and purchase history on their site to recommend other items we may be interested in buying. In this context, the party is using pattern matching concepts for a purpose which is, in some manners, mutually beneficial. It gives the company more opportunity to sell items, and may genuinely find something they would like to purchase.

Hypothetically, these same data sets could be used to make the user pay more for an item than they should. Lindell and Pinkas discuss one such scenario in "Secure Multiparty Computation for Privacy-Preserving Data Mining." One of these sites could potentially infer from such attributes as search logs, purchase history, and also duration spent on the site before purchasing, that the user does not often take time to find the best deal on an item. Knowing this, that company could possible charge that customer more for certain items. The company having more information about the customer gives it an advantage over the customer. The company knows what the customer will pay for an item, but the customer does not know how much company is willing to sell it. We could call this price discrimination.

While the previous example presents a scenario which an individual is treated unfairly with data that the company did have the right to use. However, the following example presents a way which a part, in this case the government, could use one's information to decrease their quality of life. Let us say, for example that the Federal

Bureau of Investigations were allowed created a massive database of all US citizens and their public and commercial records. Now, due to the pattern matching process, the inference is made that an individual could potentially be a terrorist. Suddenly, the innocent individual finds themselves under investigation and on the "No Fly" list.

While the scenario is fictional, it is also feasible to think that with enough data, this could be done. However, this idea of such a mistake occurring is just as possible as well. Two criticisms of the concept of using patterns matching for the means of investigation, counter-terrorism for example, are presented in "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data" by K. A. Taipale. The first issue is that the "dirty data," or records with errors and obsolete information, will lead to mistakes. The second issue, which the previously mentioned scenario fall under, would be the occurrence of false positives in the pattern matching process. We will address these issues later.

Another folly in our scenario is database itself. The Computer Matching and Privacy Protection Act of 1988 prevent government agencies from exchanging data records unless following proper protocol. Even collecting matters of public record into one database in frowned upon. Taipale makes mentions *Department of Justice v. Reporters Committee for Freedom of Press*. During these proceedings, Justice John Paul Stevens stated the following:

"Plainly there is a vast difference between the public records that might be found after a diligent search of courthouse files, county archives, and local police stations throughout the country and a computerized summary located in a single clearinghouse of information."

So, it seems that the issue is not having access to public data, but having the ability to query such data readily that is thought to be unethical. That is not to say that such large

scale databases have undertaken before. However these attempts are always met with a great deal of public outcry, regardless of its purpose. One recent case, discussed in "Privacy and Confidentiality in an e-Commerce World: Data Mining, Data Warehousing, Matching and Disclosure Limitation" by Stephen E. Fienberg, was the *Multi-state Anti-terrorism Information Exchange* system, or MATRIX, which was in operations during. In Fienberg work, it is stated that the purpose of MATRIX was to provide "the capability to store, analyze, and exchange sensitive terrorism-related information in MATRIX data bases among agencies, within a state, among states, and between state and federal agencies."

This program, developed in response to September 11th, appears to be in direct conflict with The Computer Matching and Privacy Protection Act of 1988. After much public outcry, MATRIX ceased to operate as a multi-state program in 2005.

With so many issues in privacy and ethics that could potential arise in data mining, one may beg to question why is it done at all. The important thing to remember when discussing such ethical questions to realize that all we are talking about is a tool. Tools are of great benefit to us. Many tools can be used as weapons also. Blaming the practice of data mining for being used improperly is like blaming a hammer for being used in an assault. Nevertheless, as hazardous as these tools could potential be, concept exist where these tools and be used safely.

One could recommend that we simple avoid using sensitive data at all cost. A situation Lindell and Pinkas shows why that is unrealistic. What if two hospitals wanted to perform joint research on their medical data? They cannot show one another raw data

as they would violate patient confidentiality. So, they need a way to prevent their medical data to each other so that one data set is indistinguishable from the other.

This could be accomplished using the concept of Privacy-Preserving Data Mining. Feinberg defines Privacy-Preserving Data Mining (or PPDM) as "data mining computations performed on the combined data sets of multiple parties without revealing each party's data to the other parties." The idea presented here is that those who analysis the data will never see either set of data. The only thing to that will be seen is the final result of the computation. The challenge of all this is formulating a way to pool the data without anyone seeing the raw data. This is studying as a form of cryptography known as secure multiparty computation.

So, in our hospital example, the data sets could be pooled as encrypted data and only decrypted either during the computations or possible after. Along with preventing either party from seeing each other's data, having the data encrypted up until computation would protect it from a potential third-party attack.

While PPDM could be applied to such problems as the hospital problem and would allow two government agencies to do joint research on their data as well, it does not really address some of our other problems, such as our individual on the "No Fly" list. Taipale expands upon this type of scenario with a less technical solution. In the case of a law enforcement agency, what not separate the intelligence branch from the enforcement branch entirely? If our agency were to set up protocol for passing our computed data to the enforcement side, possibly through a third party of analyst, there would be more chances that mistakes such as our false positive case could be avoided.

Even with the check against false positives and the PPDM system in place, the scenario still has the issue with the public. The agency is still in possession of a database of full of information about people. Now, let us assume that this project will continue regardless of public outcry. The agency must protect the data is now possesses from external and internal threats. One could pose the question "Who has a check on the analyst?" The question of what stops the analyst from using the database for their own malicious purposes. For this problem, we could use the concept of "selective reason." Feinberg offers the following explain from a Total Information Awareness (a predecessor to MATRIX) privacy report:

"Selective revelation works by putting a security barrier between the private data and the analyst, and controlling what information can flow across that barrier to the analyst. The analyst injects a query that uses the private data to determine a result, which is a high-level sanitized description of the query result. That result must not leak any private information to the analyst. Selective revelation must accommodate multiple data sources, all of which lie behind the (conceptual) security barrier. Private information is not made available directly to the analyst, but only through the security barrier."

Here we have a construct that prevent detailed information on an individual, thereby placing a check on the analyst as well. The analyst can only view so much information, therefore making it more difficult use private data for themselves or spoof a result. Now, the agency has access to numerous constructs to protect our information. It is now the task of the agency to use them all. It is important to understand the large degree of accountability that should be imposed if our theory database were to be created. Things such as access control to sensitive data and log records would be other concepts that are vital to our situation. One has to remember American citizens are protected from investigation without reasonable suspicion under the Amendment 4 of the US

Constitution. No matter what goal this computation is for, it is essential that all is done that can be done to ensure the integrity of the result.

This all, however, does come with a suspension of ethics. To paraphrase Pitts from his aforementioned article, the act of combining several databases together and using pattern matching as just cause for an investigation is an "an intrusive warrantless search" in itself. Even with all the safeguards discussed before, one still has to question whether this is any more ethical than online shopping company raising prices for someone just because they do not shop around.

To conclude, we could relate the practice of Data Mining to that of using a chainsaw. Both Data Mining and a chainsaw are powerful tools. Both can be a great benefit if placed in the proper hands. Both, however, can be a destructive force is placed in the wrong hands. Dangerous as these tools may be, the use of them should not be abandoned. Rather, the use of these tools should be undertaken with a great degree of responsibility and accountability while following established guidelines for their use. Finally, their misuses most come with consequences. Think of violating one's private information as a more subtle way of using the chainsaw to remove a wall to see into their home.

Works Cited

Fienberg, Stephen E. "Privacy and Confidentiality in an e-Commerce World: Data Mining, Data Warehousing, Matching and Disclosure Limitation." *Statistical Science* 21.2 (2006): 143 - 154. Web. 01 Dec 2009. <http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.ss/1154979817>.

Lindell, Yehuda, and Pinkas Benny. "Secure Multiparty Computation for Privacy-Preserving Data Mining." *Journal of Privacy and Confidentiality* 1.1 (2009): 59 - 68. Web. 01 Dec 2009. <http://jpc.stat.cmu.edu/journal/2009/vol01/issue01/lindell.pdf>.

Pitt, Joe W. "The End of Illegal Domestic Spying? Don't Count on It." *The Washington Spectator Online Edition.* (2007). 01 Dec 2009. <http://www.washingtonspectator.org/articles/20070315surveillance_3.cfm>.

Taipale, K. A. "Date Mining and Domestic Security: Connecting the Dots to Make Sense of Data". *The Columbia Science and Technology Law Review*. (2003). Web. 01 Dec 2009. <http://www.stlr.org/html/volume5/taipaleintro.php>

"The Constitution of the United States," Amendment 5.

"US CODE: Title 5,552a. Records maintained on individuals." <http://www.law.cornell.edu/uscode/05/usc_sec_05_00000552---a000-.html>

US Department of Justice v. Reporters Committee. No. 87-1379. Supreme Ct. of the US. 22 March 1989