Ethical Issues in the Field of Data Mining

CITS3200 Professional Computing

Michael Martis, 20930496

August 30th, 2013

No person can attain true privacy - participation in society itself necessitates the transfer of information, personal and otherwise, between community members (Vedder 1999). At no time during human history has this fact been more true. The widespread adoption, and increasing power, of computing technologies has irrevocably changed the way in which information is shared and collected (Charnes 2012). Billions of global citizens willingly bare personal information to their extended social network, while superbureaus maintain vast databases of indexed customer information. In this climate, data has emerged as a new form of currency; civilians

> "withdrawing cash from ATMs; paying with debit or credit cards, ... renting a car or a video; making a telephone call or an insurance claim; and, increasingly, sending or receiving e-mail and surfing the Net"
>
> (Australian Law Reform Commission 2013)

may be unaware that their personal information can be collected, traded and sold by interested third parties.

This flood of newly available data has birthed multiple technological disciplines, including that of data mining (KDnuggets 2011). The Australian Law Reform Commission (2013) defines data mining as "the large scale comparison of records or files ... collected or held for different purposes, with a view to identifying matters of interest". There are clearly ethical issues surrounding the collection and examination of personal data and, as is often the case with emerging technologies, legal systems may not yet provide adequate guidance regarding this process (Payne and Landry 2012, p 37). In the face of such issues, ethical frameworks such as the Australian Computing Society Code of Ethics (and its sister document, the Code of Professional Conduct) may be used to establish a standard of acceptable behaviour. Data mining practices that adhere to this standard can be of great benefit to society, without compromising the privacy, autonomy and equity that all individuals deserve.

As data mining is a process that inherently deals with the collection and analysis of vast quantities of potentially sensitive data, a natural concern arises for the privacy of those who

volunteer, or unknowingly yield, their personal information. The importance of such privacy has long been recognised; Article 17 of the United Nations International Covenant on Civil and Political Rights (1966) specifies that "no one shall be subjected to arbitrary or unlawful interference with his privacy" and that all people should enjoy the "protection of the law against such interference or attacks". The Australian Computer Society expresses similar concerns in their Code of Professional Conduct (2012). Section 1.2.1g explicitly outlines its members responsibility "to preserve the confidentiality and privacy of the information of others". It is imperative, then, to ensure that all data mining processes respect subject confidentiality.

It can be argued, however, that privacy is a personal issue; what qualifies as sensitive information varies widely between differing cultures and individuals (Wahlstrom et al. 2006, p 2). Furthermore, relying on a tacit assumption about the kinds of data that may be collected will almost certainly lead to mission creep (wherein the scope of the mining process grows over time) and may well misrepresent, to end users, the nature of the software with which they interact (Sultan 2012). This kind of behaviour is dishonest, and is prohibited by the Code of Professional Conduct (2012). Specifically, section 1.2.3c details a member's responsibility not to "knowingly mislead a client or potential client as to the suitability of a product or service". Hence, the safest course of action is to mandate that all data mining operations exercise transparency when it comes to the types of data collected, and the intended use of this data. This should include identifying those who will have access to collected data, how long the data will be stored, and how anonymised (by a process of aggregation or data scrambling, for example) the stored data will be. By giving individual users information about the collection of their personal data, an organisation attempts to "explicitly consider [the] interests" of those "impacted by [their] work" (ACS Professional Standards Board 2012, section 1.2.1a).

Arguably, the strong wording of the Code of Professional Conduct may warrant a more explicit approach - requiring that all data mining projects use an "opt-in" system, for example, or ensuring that data collection policies are clearly displayed before a user may interact with a system. A real-life example of such an approach can be found in the "EU Cookie Law" - a

law that requires websites to obtain consent from visitors before storing information on their computers. This approach, as opposed to hiding cookie policies in a "Terms and Services" document, appears designed to "increase the feelings of personal satisfaction ... and control" in its end users (ACS Professional Standards Board 2012, section 1.2.2d). In light of these directives, not disclosing the extent to which data is collected, or for what purpose, appears to be an unethical practice.

Simply disclosing this information, however, is not always sufficient to avoid violating the Code of Professional Conduct. This is due, in part, to the inherently exploratory nature of the data mining process. It is often impossible to know which trends, objectionable or otherwise, will emerge from a set of data until after a full analysis has been performed (Wahlstrom et al. 2006, p 2). Hence, it is difficult for an organisation to make guarantees about the type, and detail, of information they will eventually possess. In accordance with the Code of Professional Conduct's stance on honesty, this fact must be made clear to all data mining subjects. Furthermore, it remains possible to compromise the privacy of an end user, despite fully informing them about all relevant data mining policies. Amartya Bhattacharjya, a former director at the marketing management brand Unica, suggests that this can occur "when consumers have enough information, but ... are not in a position to negotiate with companies" (Shermach 2006).

For instance, a citizen may be compelled to use a piece of software in order to remain professionally competitive, despite knowing that their privacy will be compromised in the process. Interestingly, the Code of Professional Conduct does not concede to companies in this matter - the fact that a fully informed citizen may choose to use one company's product over another's does not allow them to utilise needlessly invasive data mining policies. Such policies would still constitute a failure to "respect and protect [a] stakeholder's proprietary interests" (ACS Professional Standards Board 2012, section 1.2.4d) and a failure to "provide products and services which match the operational and financial needs of [a] stakeholder" (ACS Professional Standards Board 2012, section 1.2.4a). Hence, the law should protect a user's rights in these

situations: providing them, for example, with the option of having their information removed from any database. It appears, then, that all but the most essential forms of data mining should be made optional and that as much control over the collection process as is feasible should be left in the hands of the end user.

Unfortunately, data mining legislation cannot afford end users such extensive control over the information collection process without first considering a significant mitigating factor: that of public interest and safety. The Code of Professional Conduct suggests that public interest includes "matters of public health, safety and the environment" (ACS Professional Standards Board 2012, section 1.2.1). It is conceivable that, in such matters, the benefits of certain data mining applications could justify a number of privacy breaches at the individual level. The Code of Professional Conduct appears to admit as much, when it notes that "the public interest takes precedence over personal, private and sectional interests" (ACS Professional Standards Board 2012, section 1.2.1). In 1996, for example, the United States Congress enacted the Health Insurance Portability and Accountability Act - a set of laws designed to protect, among other things, a patient's right to keep their health care information private. Although the laws contain provisions allowing for medical research, a number of researchers maintain that the act prevents necessary studies from occurring, and is hence detrimental to the health of the general public (Steinberg, Rubin, and Academic Health Centers (U.S.) 2009). Any drafted data mining legislation must account for situations in which there are conflicts between the interests of individuals and society at large.

Although the Code of Professional Conduct states that all issues "should be resolved in favour of the public interest" (ACS Professional Standards Board 2012, section 1.2.1), a practical response must be more nuanced than allowing publicly beneficial systems unrestricted access to private information. For example, data mining on crime databases, while intended to promote public safety, can lead to discrimination (Hajian, Domingo-Ferrer, and Martinez-Balleste 2011) and should not necessarily be practised in an unrestricted form. The Code of Professional Conduct accounts (2012) for such cases by specifying that its members are "expected to take

into account the spirit of [the] Code" when treating "ambiguous or contentious issues". Exactly how much public benefit is needed to justify an invasion of personal privacy is a matter of opinion, and requires the input of both domain experts, and the wider public. The Code of Professional Conduct ensures that such discussions occur by mandating that ACS members "take into consideration the fact that [their] profession traverses many other professions, and has implications for other social systems and organisations" (ACS Professional Standards Board 2012, section 1.2.1d).

Aside from privacy concerns, there are a number of other ethical issues surrounding the widespread use and influence of data mining technologies. One such issue pertains to the accuracy of information collected during the data mining process. The OECD Guidlines on the Protection of Privacy and Transborder Flows of Personal Data (2002) lists "the storage of inaccurate personal data" as a "violation of fundamental human rights". The same document mandates that any personal data mined "be accurate, complete and kept up-to-date". There are technical reasons to doubt the reliability of information stored in most personal knowledge databases. Personal data is often mined from a vast number of sources, each of varying quality, and with very few guarantees of validity or up-to-datedness (Cavoukian 1997). Inaccurate personal data becomes an ethical concern when its reliability is assumed by external parties. For example, mined data could erroneously suggest that a subject has had a past criminal conviction, which would certainly damage their merit in the eyes of prospective employers.

Clearly, allowing assumptions of this nature to routinely occur would "breach public trust in the profession" (ACS Professional Standards Board 2012, section 1.2.3), and would constitute a failure to "consider [the public's] interests" (ACS Professional Standards Board 2012, section 1.2.1a). Aside from warning end users and clients against blindly trusting information from personal knowledge databases, mandatory data scrubbing and cross-referencing procedures could be introduced to ensure that any errors therein are identified and corrected in a timely manner.

A related, but distinct, ethical issue arises when perceived trends in mined data are used to select individuals for preferential, or discriminatory, treatment. Responses or requests found to

be common to a social, racial, interest or age group can be identified, and used to inform future interactions with members of the group (Rygielski, Wang, and Yen 2002, p 488). This technique is widely used to power some of the most convenient and appreciated applications of data mining: intelligent product recommendation, as seen in websites like Amazon.com, provides a quintessential example (Corbo 2013). Crime data mining - the process of identifying trends in the personal information of convicted criminals - proves a more worrying application. Many researchers have noted that crime data inference can leave subjects in danger of stereotyping: citizens from "high risk" geographic, social, economic and racial groups could suffer unprovoked, disproportionate scrutiny and prejudicial treatment from law enforcement agencies (Hajian, Domingo-Ferrer, and Martinez-Balleste 2011).

When important decisions are made automatically, and on the basis of inferred data trends, concerns surface regarding the fair treatment of end users. These concerns are multifaceted. Firstly, data mining is an inherently extrapolative process. As such, any trends found in data may be merely perceived – rather than truly accurate – or based on skewed sampling (Nisbet, Elder IV, and Miner 2009). However, even if accurate trends could be reliably identified, there would still exist a fundamental ethical issue with their use in automated decision making. A trend can only reflect facts held in the current database - it cannot claim to make character judgements about an individual from the relevant interest group. If perceived trends are given too much weight, end users will be

> "dealt with on the basis of the attributes of the group to which they (in many cases, by chance) belong rather than on the basis of their own particular characteristics and merits" (Wahlstrom et al. 2006, p 5)

This obviously constitutes discrimination against such users, and is hence an unethical practice. This analysis is supported by the Code of Professional Conduct, which notes that

> "opportunities for employment, advancement, remuneration and other working conditions [should be] based on the actual skills and performance of employees, free of

stereotypes and prejudices."

(ACS Professional Standards Board 2012, section 1.2.6c)

To protect against discrimination of this kind, laws should be passed to limit the possible scope of special treatment based on informational trends. Furthermore, there should exist clear legal recourse for those who feel that they have been unfairly disadvantaged by such policies, and a watchdog agency should be introduced to identify discriminatory dealings of this kind.

One last area of concern, which will not be treated in this essay due to space constraints, is that of data security. The widespread practice of data mining has lead to the proliferation of personal data warehouses (Agarwal, Singh, and Pandey 2010), which can become targets for criminal activity. The law should mandate that all reasonable measures (encryption, for example) be taken to protect the information held therein from hackers and cyber-criminals.

The Australian Computer Society notes that information technology "has been beneficial to a very great extent", but "has also had some negative effects, and will continue to do so" (ACS Professional Standards Board 2012). This sentiment equally applies to the practice of data mining. While there are clearly ethical issues associated with the data mining process, none of them appear insurmountable. The discussion above attempts to show that data mining can be practised in a manner that respects the public interest, individual rights, and personal preferences. Furthermore, the power and applicability of such technologies is too great to ignore; with the right legal and ethical guidelines in place, data mining can become a tool "to enhance the quality of life of people" (ACS Professional Standards Board 2012, section 1.2.2a), and an essential, beneficial part of tomorrow's society.

References

ACS Professional Standards Board (July 2012). *ACS Code of Professional Conduct*.

Agarwal, Sonali, Neera Singh, and GN Pandey (2010). "Implementation of Data Mining and Data Warehousing In E-Governance." In: *International Journal of Computer Applications* 9.4, pp. 18–22.

Australian Law Reform Commission (January 2013). *Overview: Impact of Developing Technology on Privacy*. URL: http://www.alrc.gov.au/publications/9.%20Overview%3A%20Impact%20of%20Developing%20Technology%20on%20Privacy/data-matching-and-data-mining (visited on 08/28/2013).

Cavoukian, Ann (1997). *Data mining: Staking a claim on your privacy*.

Charnes, John (June 2012). *How has technology changed the way we interact with data?* URL: http://www.syntelli.com/technology-changed-interact-data/ (visited on 08/28/2013).

Corbo, Ruben (February 2013). *Overview Of Product Recommendation Engines*. URL: http://online-behavior.com/targeting/recommendation-engines (visited on 08/28/2013).

Hajian, S., J. Domingo-Ferrer, and A. Martinez-Balleste (2011). "Discrimination prevention in data mining for intrusion and crime detection." In: *Computational Intelligence in Cyber Security (CICS), 2011 IEEE Symposium on*, pp. 47–54. DOI: 10.1109/CICYBS.2011.5949405.

KDnuggets (2011). *Introduction to Data Mining, Notes*. URL: http://www.kdnuggets.com/data_mining_course/x1-intro-to-data-mining-notes.html (visited on 08/28/2013).

Nisbet, Robert, John Elder IV, and Gary Miner (2009). *Handbook of statistical analysis and data mining applications*. Access Online via Elsevier. Chap. 20.

OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (2002). Paris: OECD Publishing. ISBN: 978-9264196391.

Payne, Dinah and Brett J. L. Landry (2012). "A Composite Strategy for the Legal and Ethical Use of Data Mining." In: *International Journal of Management, Knowledge and Learning* 1.1, p. 37. URL: http://ideas.repec.org/a/isv/jouijm/v1y2012i1p27-43.html.

Rygielski, Chris, Jyun-Cheng Wang, and David C Yen (2002). "Data mining techniques for customer relationship management." In: *Technology in society* 24.4, pp. 483–502.

Shermach, Kelly (August 2006). *Data Mining: Where Legality and Ethics Rarely Meet*. (Visited on 08/26/2013).

Steinberg, M.J., E.R. Rubin, and Association of Academic Health Centers (U.S.) (2009). *The HIPAA Privacy Rule: Lacks Patient Benefit, Impedes Research Growth*. Association of Academic Health Centers. URL: http://books.google.com.au/books?id=GGIvYAAACAAJ.

Sultan, Aisha (September 2012). *Digital data mining spurs efforts to curb use without permission*. URL: http://www.stltoday.com/news/local/metro/digital-data-mining-spurs-efforts-to-curb-use-without-permission/article_8bcb85cf-b57e-5e6d-ac61-f500c71685c4.html (visited on 08/28/2013).

United Nations (1966). *United Nations International Covenant on Civil and Political Rights*. URL: http://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx (visited on 08/26/2013).

Vedder, Anton (1999). "KDD: The challenge to individualism." English. In: *Ethics and Information Technology* 1.4, pp. 275–281. ISSN: 1388-1957. DOI: 10.1023/A:1010016102284. URL: http://dx.doi.org/10.1023/A%3A1010016102284.

Wahlstrom, Kirsten et al. (February 2006). *On the Ethical and Legal Implications of Data Mining*. Tech. rep. School of Informatics and Engineering, Flinders University.