Ethics and Information Technology **1:** 137–145, 1999.

Informational privacy, data mining, and the Internet

Herman T. Tavani
Department of Philosophy, Rivier College, Nashua, New Hamphsire, USA. E-mail: htavani@rivier.edu

Abstract. Privacy concerns involving data mining are examined in terms of four questions: (1) What exactly *is* data mining? (2) How does data mining raise concerns for personal privacy? (3) How do privacy concerns raised by data mining differ from those concerns introduced by 'traditional' information-retrieval techniques in computer databases? (4) How do privacy concerns raised by mining personal data from the Internet differ from those concerns introduced by mining such data from 'data warehouses?' It is argued that the practice of using data-mining techniques, whether on the Internet or in data warehouses, to gain information about persons raises privacy concerns that (a) go beyond concerns introduced in traditional information-retrieval techniques in computer databases and (b) are not covered by present data-protection guidelines and privacy laws.

What exactly is data mining?

While the term 'data mining' is relatively new, much of the technology used in the data-mining process is not. For example, many of the algorithms currently used in data mining are the result of research in artificial intelligence in the 1980s. Essentially, data-mining technology combines artificial intelligence, statistical analysis, knowledge acquisition from expert systems, data visualization, machine discovery, and pattern recognition. Cavoukian defines data mining as a 'set of automated techniques used to extract buried or previously unknown pieces of information from large databases.'[1] And Bigus notes that data mining can be viewed as a technique for the 'efficient discovery of valuable, nonobvious information.'[2] Using data-mining techniques it is possible to unearth patterns and relationships, which were previously unknown, and to use this 'new' information, i.e., new facts and relationships in the data to make decisions and forecasts.

In its broadest sense, data mining is the process of (A) finding patterns or correlations in the data (e.g., records) stored in large databases and (B) analyzing that data from different perspectives, categorizing it, and summarizing it into useful information. Some computer scientists further distinguish between data mining and the knowledge discovery in databases (KDD) process. Whereas data mining is the process for *discovering* patterns in data, KDD includes the work done before the data is searched for patterns (e.g., processes such as 'data preparation,' 'data selection,' and 'data cleansing') as well as the work done on the patterns after searching (i.e., the "incorporation of appropriate knowledge" and "proper interpretation" of the data) so as to make the data useful. Fayyad, Piatetsky-Shapiro, and Smyth use the term 'KDD' to refer to the overall process of "discovering useful knowledge from data," while they regard data mining as a particular step in this process, viz., 'the application of specific algorithms for extracting patterns of data.'[3] Whereas data mining per se' involves 'determining patterns' from the data, the additional steps in the KDD process ensure that 'useful knowledge' is derived from that data. For purposes of this study, however, the expression 'data mining' is used in its generic sense to refer to the overall process of preparing data, discovering patterns in data, and analyzing that data into useful knowledge.

How does data mining raise concerns for personal privacy?

Privacy, which is often associated with, and sometimes described in terms of, liberty, autonomy, solitude, and secrecy, is a concept that is not easily defined. Moor points out that in the United States the concept of privacy has evolved from one concerned primarily

[1] A. Cavoukian. *Data Mining: Staking a Claim on Your Privacy*. Information and Privacy Commissioner's Report, Ontario, Canada, 1998.

[2] J.P. Bigus. *Data Mining With Neural Networks*. McGraw-Hill, New York, 1996.

[3] U. Fayyad, G. Piatetsky-Shapiro and Smyth Padhraic. The KDD Process for Extracting Useful Knowledge From Volumes of Data. *Communications of the ACM* (November), 39(11): 27–34, 1996.

with intrusion into one's personal space and interference with one's personal affairs to one currently concerned primarily with personal information and access to personal information.[4] Some privacy analysts now speak of 'informational privacy' as a category of privacy with a set of issues that are distinguishable from privacy concerns related to intrusion and interference, which are sometimes described as 'psychological privacy'[5] or 'associative privacy'.[6] We shall see that privacy concerns arising from data mining fall primarily under the category of informational privacy.

We often hear remarks to the effect that one's privacy has been 'lost,' 'diminished,' 'intruded upon,' 'invaded,' 'violated,' 'breached,' and so forth. Each of these descriptions, in turn, reflects the insights and biases of one or more models or theories of privacy. For example, some theories see privacy as an 'all-or-nothing' concept, i.e., privacy is something that one either has (totally) or does not have. Other theories view privacy as something that can be diminished, e.g., as a repository of information possessed by an individual, which can be eroded gradually. Still others see privacy in terms of a spatial metaphor such as a zone that can be intruded upon or invaded by others. And other theories view privacy in terms of confidentiality that can be violated, or trust that can be breached. There are several interesting theories or models of privacy that either directly correspond to, or approximately fit, one or more of the metaphors described above. To examine these various theories would, however, take us beyond the scope of the present study.[7] For purposes of the present study, our analysis will be centered on whether data mining "raises concerns" for privacy, or more specifically concerns for informational privacy.

With respect to informational privacy, there are at least two important ways in which the introduction of a new technology can raise privacy concerns: (i) the technology is used to collect information about an individual or group of individuals without the awareness or knowledge of the individual(s) about whom the information is being collected; and (ii) individuals are aware that information about them is being collected via a certain technology but have no have no say in how the information about them is used (disclosed, exchanged, sold, etc.). We shall see that in the case of data mining, information is typically collected about individuals without the awareness or knowledge of those individuals. And we shall see that even when individuals are aware that information about them is being collected, certain controversies still arise because those individuals cannot possibly be told in advance what kind of information data-mining algorithms will yield about them and how that information will be used. Data-mining programs, by their very design, reveal information about individuals that would have been extremely difficult for data users (those who use data mining to collect information) to foresee and for data subjects (those about whom the data is collected) to consent.

A technology can also raise concerns for privacy when the information about persons being collected is not covered by provisions in current privacy laws, such as the U.S. Privacy Act of 1974, or in current data-protection guidelines, such as the Fair Information Practices (FIPs). FIPs are codified in the OECD (Organization for Economic Cooperation Development) principles, which include eight internationally agreed upon principles related to the collection, use, and disclosure of information about persons. Two of those principles have to do with *specifying the purpose* and *limiting the use* of information on data subjects (individuals and groups) by data users (such as businesses and governments). We will see that data-mining techniques, by their very nature, cannot comply with these two principles, and thus are incompatible with current data-protection guidelines. We will also see how privacy concerns raised by data mining go beyond those covered in privacy laws such as the Privacy Act of 1974 in the U.S.

Privacy concerns raised by data mining vs. traditional information-retrieval techniques

We can ask whether privacy concerns raised by data mining differ in any meaningful respects from those concerns introduced in 'traditional' practices of retrieving personal information from computer databases. Such traditional practices include *computer merging* (i.e., the merging of electronic records across computer databases) and *computer matching* (or the matching of electronic records against databases).[8] On

[4] J.H. Moore. Towards a Theory of Privacy in the Information Age. *Computers and Society* (September), 27(3): 27–32, 1997; Reason, Relativity, and Responsibility in Computer Ethics. *Computers and Society* (March), 28(1): 14–21, 1998.

[5] P.M. Regan. *Legislating Privacy: Technology, Social Values, and Public Policy*. University of North Carolina Press, Chapel Hill, NC, 1995.

[6] J.W. DeCew. *In Pursuit of Privacy: Law, Ethics, and the Rise of Technology*. Cornell University Press, Ithaca, 1997.

[7] Normative theories of privacy and their implications for data mining are considered in a separate paper, titled "Data Mining, Personal Privacy, and Public Policy," which I presented at the *CEPE '98 Conference* (London School of Economics), December 14, 1998. See, *Proceedings of the CEPE '98 Conference* L. D. Introna, editor, 113–120.

[8] For a discussion of computer merging and computer matching, see Tavani, 1996.

the one hand, privacy concerns associated with data mining would seem to have a great deal in common with traditional computerized techniques used in the collection, retention, and exchange of personal information. After all, both techniques depend on the use of large computer databases to record, store, and exchange personal information. Although privacy concerns raised by data mining may share many similarities with privacy concerns raised by traditional database retrieval techniques, such as those involved in the merging and matching of computerized records, there are at least six ways in which privacy concerns raised by data mining go beyond concerns resulting from traditional information-retrieval techniques in computer databases:

(1) the *implicit* patterns involving information about persons that can be derived form data in the data-mining process vs. the explicit nature of the personal data (in records) extracted in traditional database retrieval techniques.
(2) the use of (possibly) a single database (or 'data warehouse') to extract information about persons vs. the use of multiple databases to exchange and retrieve such information.
(3) the use of 'open-ended' queries to *discover* information on relationships and associations about individuals and groups of individuals vs. (traditional) specific queries to retrieve information about relationships and associations that are already known to exist.
(4) the nonpredictive aspect of information about persons gained from data mining vs. the generally predictive aspect of information retrieved from traditional database techniques.
(5) the *public* nature of much of the information about persons that is extracted through the data-mining process vs. the private or intimate nature of the information about persons retrieved and exchanged in traditional database-exchange techniques.
(6) the ability to construct new groups or categories of persons based on patterns of information derived from data mining vs. the mere extraction of information about individuals themselves from personal data accessible to traditional techniques of database retrieval.

Let us next consider each of these distinctions in more detail in order to see how they raise concerns for personal privacy that go beyond those concerns introduced by traditional information-retrieval techniques such as computer merging and computer matching.

Why are the six distinctions relevant?

First, in data mining the information about persons extracted from a database is not necessarily explicit in the records contained in the database. Instead, implicit patterns and associations are *discovered* among the data that reside in the database. Such is not the case, however, in computer merging and computer matching techniques. When computer matches are performed, for example, the identities of specific records are used or requested, i.e., records with particular identifiers such as an individual's name, ID number, and so forth, already be explicit in the database. And in traditional techniques involving the merging of computer records, only explicit records, or data contained in specific fields of records, about individuals are used to create a merged file.

Second, whereas the merging or integrating of electronic records across computer databases and the matching of electronic records against databases both involve the *exchange* of (explicit) records involving more than one database, the data-mining process involves the search for implicit patterns and associations in data that *can* reside in only one large database or in what is commonly referred to as a 'data warehouse' (see Section 4 of this study for more detail). So the use of (potentially) a single database for extracting personal information is yet another feature that distinguishes the data-mining process from traditional information-retrieval techniques. That is, in the data-mining process it is not essential that data (for example, records in a "data warehouse") be transferred to, or exchanged with records in, an external database. For instance, WalMart, a retail chain in the U.S., mines information about its customers from a single database, viz., its own proprietary data warehouse.

Third, in traditional information-retrieval practices, database records (or tuples derived from fields of records) are returned in response to a specific query, e.g., a query about the identity of a specific name, ID, etc. Data-mining software, on the other hand, can be used to extract personal information about relationships and patterns in data, based on "open-ended" user inquiries. Traditional database queries entered in business databases can answer questions like, "How many widgets did our company sell in the UK in 1997?" The relationships that exist among these data are already known, in some sense, to the user, who, by framing the proper question, e.g., "how many *X's* were purchased by *Y*?" obtains the desired answer. Data mining, however, uses "discovery-based" approaches in which pattern-matching and other algorithms are used to *discover* key relationships in the data, which were previously unknown to the user. The discovery model is different because the system automatically discovers

information "hidden" in the data, i.e., in open-ended queries the data is 'sifted' in search of frequently occurring patterns, trends, and generalizations about the data without intervention or guidance from the user. For example, the user could simply conduct a query with a request or command such as "show all patterns" or "show a category of trends/relationships." Before data-mining techniques were employed in large databases, individuals might have had a false sense of comfort regarding personal information about them, believing that there was possibly too much data to be analyzed intelligently. Data-mining software, however, now makes it possible for terabytes of data containing personal information to be examined for meaningful patterns.

The detection of patterns and forecasting data, so easily derived from open-ended queries in the use of data-mining software, is closely related to our fourth point of distinction, viz., that companies who practice data mining cannot always *predict* what uses the resulting information will have. Again, this is not typically the case with information gained from traditional practices of information retrieval. Law enforcement agencies that engage in record matching can predict the likely outcome of matched records, e.g., 'hits' identifying the names of individuals whose electronic records reside in two or more databases. Traditional computer-merging practices also have a predictive aspect as well, since they are often used intentionally to construct a composite picture or mosaic of one or more individuals, based on specific records about that individual that reside in more than one database. However, since data mining is based on the extractions of unknown patterns information from a database, users of data-mining programs cannot predict, i.e., they cannot know at the outset what kind of potentially valuable personal data or what kinds of relationships among the data will emerge.

Next, there is the distinction between the public vs. the private (or intimate) nature of information typically mined. A considerable amount of the information mined about individuals comes from data gathered in the public as opposed to the "intimate" sphere. Nissenbaum characterizes much of the gathering of personal information in public transactions by vendors in the commercial sector, which would also seem to include personal information mined by businesses, as 'public surveillance'.[9] Under this category, she includes the 'collection, collation, and transmission of information,' even though much of the 'surveillance' itself occurs in the nonintimate realm. (For discussions of surveillance in what some now call a 'panopticon society,' see Gandy, van den Hoven, and Blanchette and Johnson.)[10] In traditional practices of exchanging information in computer databases, especially in computer-merging techniques, the primary kind of information exchanged about persons has been 'confidential' information, such as individual's financial or medical records.

Finally, many data-mining practices result in information gained about a certain group of individuals rather than information about the individuals themselves. Consider, for example, a bank database which is mined to discover the possible groups of customers it can target for various mailing campaigns. In such a case, the data is searched with no hypothesis in mind other than for the data-mining algorithm to group the customers according to the common characteristic found. We might ask how exactly this process is different from traditional information-exchanging techniques. In computer matching, electronic records involving a category or class of individuals (e.g., government employees) in one database have been matched against a database containing records about another group or class of individuals (e.g., welfare recipients) in the expectation that certain matches or 'hits' would result. In matching, the aim is to discover information about particular individuals who happen to be members of pre-selected categories or groups, not about information regarding the groups themselves. Such is not the case with data mining, however, where pattern-matching algorithms are run to extract information about groups of individuals and patterns within groups.

A hypothetical scenario

To illustrate some of the key points discussed in the preceding sections, consider the case of Lee, a junior executive at the ABC Marketing Firm in the U.S. Lee applies for an automobile loan at a local bank. To secure the loan for the purchase of a new automobile, Lee agrees to complete the usual forms required by the bank for loan transactions. For example, Lee indicates that he has been employed at the ABC company for more than three years and that his current annual

[9] H. Nissenbaum. Can We Protect Privacy in Public? *Proceedings of the Conference on Computer Ethics: Philosophical Enquiry* (*CEPE '97*), 191–204. Erasmus University, Rotterdam, the Netherlands, 1997.

[10] O.H. Gandy. *The Panoptic Sort: A Political Economy of Personal Information*. Westview Press, Boulder, CO, 1993; J. Hoven van den. Privacy and the Varieties of Moral Wrong-Doing in the Information Age. *Computers and Society* (September), 27(3): 33–37, 1997; J.-F. Blanchette and D.G. Johnson. Cryptography, Data Retention, and the Panopticon Society. *Proceedings of the Conference on Computer Ethics: Philosophical Enquiry* (*CEPE '98*). The University of London Press, London, UK, 94–105, 1998.

salary is $90,000. He also indicates that he has $10,000 dollars in a separate savings account which he intends to use as a down-payment for the purchase of a new BMW. On the loan form, Lee also indicates that he is currently repaying a $15,000 dollar personal loan used to finance a family vacation to Europe taken during the previous year.

Thus far, the transaction between Lee and the bank would seem quite appropriate in that Lee wishes to borrow money from the bank, and the bank would seem to have a legitimate need to get appropriate information about Lee to make an informed decision as to whether or not to grant Lee the loan. To acquire the loan, Lee has authorized the bank to have information about him, i.e., information about his current employment, salary, savings, outstanding loans, etc.

While Lee has given the bank information about himself for use in one context, viz., to make a decision about whether or not he should be granted a loan to purchase a new automobile, Lee should also be able to expect that the information given to the bank will not be exchanged with a third party (or at least not without Lee's knowledge and consent). And while the bank has agreed not to exchange or disclose information about Lee to a third party, it is unclear whether the bank has agreed not to use the information it now has about Lee for certain in-house analyses.

Next, suppose that the bank mines information from its databases and discovers the following pattern: Executives earning more than $70,000 but less than $120,000 annually, and who purchase luxury cars (such as BMWs), and who take expensive vacations, often go into business for themselves within five years of employment. A separate pattern-matching program reveals that the majority of marketing entrepreneurs who go into business for themselves declare bankruptcy within one year of starting their own businesses. All of a sudden, Lee is a member of a group that neither he nor possibly even the loan officers at the bank had ever known to exist, viz., the group of marketing executives likely to start a business and declare bankruptcy within a year of starting such a business. With this new category and with this 'new information' about Lee, the bank determines that Lee, and people that fit into Lee's group, are long-term credit risks.

Why does the mining of data about Lee by the bank raises concerns for privacy. While Lee voluntarily gave the bank information about his annual salary, about previous loans involving vacations, and about the type of automobile he intended to purchase, he gave each piece of information for a specific purpose and use. Individually, each piece of information was appropriately given in order that the bank could make a meaningful determination about Lee's request for an automobile loan. However, it is by no means clear that Lee authorized the bank to use disparate pieces of that information for more general data-mining analyses that would reveal patterns involving Lee that neither he nor the bank could have anticipated at the outset.

Let us next consider ways in which the case involving Lee illustrates the six characteristics associated with data mining. First, the information about Lee's being someone likely to start his own business, which would ultimately lead to his declaring personal bankruptcy, was not explicit in any of the data (records) about Lee; rather it was implicit in patterns of data about people similar to Lee in certain respects but vastly different from Lee in other important respects. Second, the information about Lee was extracted from one or more databases internal to the bank and was not transferred to or exchanged with one or more external databases. Third, the information about Lee was discovered via an open-ended query and not through a specific query about Lee himself. Fourth, there was no way the bank could have predicted what kinds of information about Lee and similar customers that would result from the execution of various pattern-matching algorithms used in the data-mining process. Fifth, at least some of the information about Lee, e.g., information that he took a vacation in Europe the previous year can be considered public rather than private or intimate information about Lee. (In the case of Lee, however, the public vs. private distinctions regarding certain information used to make the loan decision are less critical than in many other cases of data mining, where more intimate or confidential information about persons can be used.) Finally, Lee's case illustrates how the data-mining process can be used to construct new categories and groups of individuals such that the persons who eventually become identified with those groups would very likely have no idea that they would be identified with such groups and would have decisions made about them by virtue of being identified as members of those groups. For example, it is somewhat doubtful that Lee would have known that he was a member of a group of professional individuals likely to start a business, and that he was a member of a group whose businesses were likely to end in bankruptcy. The discovery of such groups are, of course, a result of data mining.

As noted in the preceding paragraph, no information about Lee was exchanged with databases outside the bank. So the bank did not transfer data about Lee to an external database without Lee's consent. However, the bank did use information about Lee internally in a way that he had not explicitly authorized. And it is in this sense of unauthorized internal use by data users that data mining raises serious concerns for personal privacy. Note also that even if Lee had been granted the loan for the automobile, serious privacy concerns

would still have been raised by the bank's data-mining practices. For Lee was merely one of many bank customers who had voluntarily given certain personal information about themselves to the bank for use in one context (say, for example, a loan request) and then had that information about them, which was authorized for use in one context, subsequently used in ways that were not specifically authorized.

Implications for current data-protection guidelines and privacy laws

What implications does the previous scenario have for our current data-protection guidelines? The Code of Fair Information Practices mentioned earlier includes a number of principles, such as those concerned with data quality, purpose specification, use limitation, openness, individual participation, etc., which were implemented in the OECD guidelines in 1980 and which have become internationally agreed upon principles. It would seem that certain data-mining practices are clearly incompatible with at least two of the OECD Principles: *Purpose Specification* and *Use Limitation*. According to the Purpose Specification Principle, the "purpose for which data are collected *should be specified no later than at the time of data collected*" (italics added). And according to the Use Limitation Principle, 'Personal data should not be disclosed, made available, or *otherwise used for purposes other than those specified with the Purpose Specification Principle* except (a) with the consent of the data subject, or (b) by the authority of law' (italics added). In the case of Lee in the preceding section, all of the purposes for which the data were collected were not specified to Lee at the time of data collection, and the information collected about Lee was used for 'purposes other than those specified in accordance with the Purpose Specification Principle' without Lee's consent.[11]

As noted in the preceding section, independent of whether Lee was eventually denied or granted the loan for the automobile, a misuse of the information collected about Lee occurred. And we saw that it was not only the information about Lee that was misused in the data-mining practices. All of the individuals who had information about them given to the bank for use in one context used by the bank in a context for which they had not given their explicit consent, viz., in the data-mining analyses, had, according to the OECD guidelines, information about them misused. So it is in this sense that the mining of personal data is incompatible with current data-protection guidelines.

Data mining would also seem to enjoy little protection from the U.S. Privacy Act of 1974. Although the Act is concerned with the fair use of personal information, it seems to address more specifically the transfer and exchange of personal data between and among databases. Although the mining of personal data can, as we have seen, be accomplished within a single database and thus does not require the exchange of personal data across multiple databases, it is important to note that the mining of such data is nonetheless incompatible with the spirit of the Privacy Act of 1974. So it would seem that in the U.S. more up-to-date privacy legislation is needed to address explicitly those privacy concerns raised by data mining.

Mining personal data from the Internet vs. 'data warehouses'

The discussion of data mining and privacy thus far has centered on cases involving the mining of personal data from large databases, sometimes referred to as *data warehouses*. These 'warehouses,' which are huge, highly integrated databases, are typically used for processing transactional information for sales and marketing. Some analysts and consumer advocacy groups are concerned that data about persons can and soon will be mined from the Internet as well. In this section, we consider whether there are any relevant differences with respect to privacy concerns related to the mining of personal data from the Internet as opposed to mining that data from data warehouses.

Cavoukian notes that although data warehouses are not an essential to the data-mining process, the mining potential of data can be significantly enhanced when the appropriate data are stored in a data warehouse.[12] Through data warehousing, the process of extracting and transforming operational data into informational data in a 'central data store' or warehouse, data can be managed from a single database. So data warehousing introduces greater efficiency in the data mining process, which has also resulted in that process becoming more economical for businesses who elect to adopt it.

Many analysts who view the data warehouse as the 'ideal structure' for data mining (see Inmon),[13] also believe that the Web, which is considerably less structured than data warehouses in those respects key to current data-mining techniques, is a 'quagmire' for

[11] A. Cavoukian also notes the incompatibility of data mining with the certain OECD guidelines. However, her account of this incompatibility with respect to specific OECD Principles, which is based in large part on her own model of privacy, differs in certain key respects from the account given here. For more details, see Cavoukian (1998).

[12] *Data Mining: Stating a Claim on Your Privacy*.

[13] W.H. Inmon. The Data Warehouse and Data Mining. *Communications of the ACM* (November), 39(11): 49–50, 1996.

mining data. Oren Etzioni and Fulda, however, believe that the Web is a potential 'gold mine' for extracting personal data.[14] And Cavoukian (1998), who points out that one of the purposes of data mining is to "map the unexplored terrain of the Internet," notes that the Internet is becoming an "emerging frontier for data mining." She notes that with access to an Internet server, it is possible to FTP (file transfer protocol) the data from the client's server and then conduct various data mining activities.

Because data-mining software employs certain AI techniques, it can "learn" about the Web by coming to understand the content associated with common HTML tags (see, for example, Fulda).[15] Eisenberg notes that intelligent agents can "sift through" the potential wealth of data on the Internet, and Etzioni describes the use of *learning techniques* or systems such as 'softbots' (intelligent software robots or agents that use tools on a person's behalf) and metasearch engines (such as MetaCrawler and Ahoy) to uncover general patterns at individual Web sites and across multiple Web sites.[16] So data-mining techniques that currently raise privacy concerns at the database (or data warehouse) level may very likely soon raise such concerns on the Internet and the Web as well.

Fulda also points out that currently most of the information on the Web about an individual who is not a public figure is there 'by his leave', e.g., his home page, items he has chosen to publish, etc.[17] Thus far, much of that information included on the Web has not yet proved to be a practical repository for those who mine personal data. If Etzioni's, Eisenberg's, and Fulda's assessments are correct, however, that may soon change. For as more and more personal information is successfully mined from Web sites, including information from home pages, individuals may become more cautious and perhaps more selective about which pieces of personal information they are willing to include in personal home pages as well as on the pages in related Web sites that they may also happen to maintain.

Why are distinctions involving mining data from the Internet in general, and the Web in particular, as opposed to mining it from data warehouses relevant for our discussion concerning privacy? In Section 3 of this study, we examined certain key differences between privacy issues associated with traditional information-exchanging practices in databases and privacy issues related to data mining. Despite genuine differences, however, the privacy concerns related to each can be described as having one factor in common, viz., both are instances of what Johnson and Nissenbaum (1995) call 'information privacy,' i.e., privacy issues related to databases. In so far as data warehouses are used as the source from which personal data is mined, privacy concerns surrounding data mining would clearly seem to be an instance of information privacy. One critical distinction between personal information extracted from a data warehouse vs. that which is extracted from the Web, however, is that in data warehouses the personal information or data extracted is "hidden" from public view, whereas much of the (nontransactional-related) personal information extracted from Web sites is, in effect, already available for public viewing.

How is this hidden vs. public distinction regarding the nature of the personal data extracted relevant? In a previous work, I argued that privacy concerns surrounding certain uses of Internet search-engine technology arise in spite of the publicly available aspect of personal information on the Internet.[18] I also argued that this distinction makes it difficult to classify privacy issues related to search engines as falling into either one of Johnson and Nissenbaum's categories of 'information privacy' or 'communications privacy', a distinction which works remarkably well as a classifying principle for most computer-related privacy issues. And since privacy issues emerging from certain uses of Internet search engines could not be examined under either of Johnson and Nissenbaum's privacy categories, it was suggested that those specific privacy concerns could not be reduced to or analyzed simply in terms of those traditional categories. Based on the preceding examples involving the mining of personal data from the Internet, the same would seem to be true for privacy concerns involving Internet-related data mining. So we may not be able to approach privacy issues related to data mining on the Internet in the same way we have analyzed more traditional computer-related privacy concerns involving databases.

We have also seen that information stored in, and retrieved from, commercial data warehouses is primarily transactional in nature, or at least in its origin. Personal information mined from the Web, however, need not be (and frequently is not) transactional. For example, information typically included in

[14] O. Etzioni. The World Wide Web: Quagmire or Gold Mine? *Communications of the ACM* (November), 39(11): 65–68, 1996; J. Fulda. Data Mining and the Web. *Computers and Society* (March), 28(1): 42–43, 1998.

[15] *Ibid.*

[16] A. Eisenberg. Privacy and Data Collection on the Net. *Scientific American* (March), p. 120, 1996; *Communications of the ACM*: 65–68.

[17] *Computers and Society*: 42–43.

[18] H.T. Tavani. Internet Search Engines and Personal Privacy. *Proceedings of the Conference on Computer Ethics: Philosophical Enquiry* (*CEPE '97*), 214–223. Erasmus University Press, Rotterdam, the Netherlands, 1997.

personal home pages and in various Web sites that are not commercial-based is nontransactional. Of course, much transactional information can now also be gained from the Web as well, because of recent trends in Internet commerce. For example, when an individual orders a book from Amazon.com (an online book store), transactional information is recorded about the purchase, and information about that particular transaction can be (and frequently is) used for future business decisions. However, what distinguishes the Internet as a potential mining resource from large commercial databases used in data mining is the vast amount of nontransactional, personal information currently available on the Web that could also be mined. Can this personal information, which is also public in some sense, be protected?

Nissenbaum notes that very few users of the Internet realize that their activities may be "placed under surveillance."[19] She goes on to note that information such as user's email addresses as well as the system and network characteristics of a user's computer are easily recorded by many of the Web sites one visits. And Eisenberg notes that mouse clicks and key strokes, or what she calls "clickstreams", are frequently recorded by owners and operators of many Web sites.[20] That is, information about which Web sites a user visits, how long he or she stays there, and where he or she goes afterward are recorded. This raw data about an individual's online behavior can then be transformed into useful information, i.e., in many cases a kind of transactional information which can used by an online businesses for future applications, exchanged with other online businesses, or sold to businesses that operate in the physical realm.

It would seem then that many of the privacy concerns regarding data mining on the Web, like those in data warehouses, are not so much involved with personal information related to confidential or intimate matters (e.g., information including one's medical records or bank records); rather, issues arise because seemingly harmless pieces of information about persons can be 'excavated' from an individual's online activities and used in a way to construct a profile of an individual based on information freely put by that individual on the Web for use in a particular context. And since that context into which the information is put might not be a business context, online businesses who use that information for business purposes would seem to engage in what Nissenbaum calls the 'violation of contextual integrity.'[21] As we saw earlier, concerns with informational privacy generally relate not to the collection of information itself, which many consumers would gladly give for appropriate use in a specific context, but to the manner in which personal information is collected, used, and then disclosed. We also saw that when a business collects information without the knowledge or consent of the individual to whom the information relates, or uses of that information in ways that are not known to the individual, or discloses the information without the consent of the individual, informational privacy is seriously threatened. And since data mining, by its very nature, makes possible such practices, the mining of personal data from the Web in particular, and the Internet in general, raises serious concerns for personal privacy.

Conclusion

It has been argued that certain data-mining techniques, whether used in data warehouses or on the Internet, to extract information about individuals raise serious concerns for personal privacy. We saw that one reason why such techniques cause privacy concerns is because individuals are often not aware that data about them which they may have authorized for collection and use in one context is being mined, in ways they had not explicitly authorized, into information that is useful to certain businesses and organizations. Even though individuals might have explicitly authorized information about themselves to be collected for use by a business in one context, it does not follow that those individuals have also authorized that such information can then be subsequently mined for further use and analysis. We also saw that because of the way the data-mining process works, businesses and organizations (data users) who engage in data-mining practices cannot possibly inform individuals or groups of individuals (data subjects) in advance as to how information collected about them in one context will be used in future information-retrieval activities, since the data users themselves are unable to know which kinds of new groups or categories will emerge. We saw that because of these and other factors, data mining is incompatible with both the Purpose Specification and Use Limitation Principles of the Fair Information Practices, an international set of guidelines codified in the OECD, as well as with the spirit of the U.S. Privacy Act of 1974. Thus, it would seem to follow that more specific data-protection guidelines and more

[19] *Proceedings of the Conference on Computer Ethics: Philosophical Enquiry*: 191–204.

[20] *Scientific American*, p. 120.

[21] *Proceedings of the Conference on Computer Ethics: Philosophical Enquiry*: 191–204.

up-to-date privacy laws, which take into account the new kinds of privacy threats posed by data mining, are needed.

References

Joseph P. Bigus. *Data Mining With Neural Networks*. McGraw-Hill, New York, 1996.

Jean-Francois Blanchette and Deborah G. Johnson. Cryptography, Data Retention, and the Panopticon Society. *Proceedings of the Conference on Computer Ethics: Philosophical Enquiry* (*CEPE '98*). The University of London Press, London, UK, 94–105, 1998.

Ann Cavoukian. *Data Mining: Staking a Claim on Your Privacy*. Information and Privacy Commissioner's Report, Ontario, Canada, 1998.

Judith W. DeCew. *In Pursuit of Privacy: Law, Ethics, and the Rise of Technology*. Cornell University Press, Ithaca, NY, 1997.

Anne Eisenberg. Privacy and Data Collection on the Net. *Scientific American* (March), p. 120, 1996.

Oren Etzioni. The World Wide Web: Quagmire or Gold Mine? *Communications of the ACM* (November), 39(11): 65–68, 1996.

Usama Fayyad, Gregory Piatetsky-Shapiro and Smyth Padhraic. The KDD Process for Extracting Useful Knowledge From Volumes of Data. *Communications of the ACM* (November), 39(11): 27–34, 1996.

Joseph Fulda. Data Mining and the Web. *Computers and Society* (March), 28(1): 42–43, 1998.

Oscar H. Gandy. *The Panoptic Sort: A Political Economy of Personal Information*. Westview Press, Boulder, CO, 1993.

Jeroen Hoven van den. Privacy and the Varieties of Moral Wrong-Doing in the Information Age. *Computers and Society* (September), 27(3): 33–37, 1997.

W.H. Inmon. The Data Warehouse and Data Mining. *Communications of the ACM* (November), 39(11): 49–50, 1996.

Deborah G. Johnson and Helen Nissenbaum, editors. *Computers, Ethics & Social Values*. Prentice Hall, Englewood Cliffs, NJ, 1995.

James H. Moor. Towards a Theory of Privacy in the Information Age. *Computers and Society* (September), 27(3): 27–32, 1997.

James H. Moor. Reason, Relativity, and Responsibility in Computer Ethics. *Computers and Society* (March), 28(1): 14–21, 1998.

Helen Nissenbaum. Can We Protect Privacy in Public? *Proceedings of the Conference on Computer Ethics: Philosophical Enquiry (CEPE '97)*, pp. 191–204. Erasmus University, Rotterdam, the Netherlands, 1997.

Priscilla M. Regan. *Legislating Privacy: Technology, Social Values, and Public Policy*. University of North Carolina Press, Chapel Hill, NC, 1995.

Herman T. Tavani. Computer Matching and Personal Privacy: Can They Be Compatible? *Proceedings of the Symposium on Computers and the Quality of Life (CQL '96)*, pp. 97–101. ACM Press, New York, 1996.

Herman T. Tavani. Internet Search Engines and Personal Privacy. *Proceedings of the Conference on Computer Ethics: Philosophical Enquiry (CEPE '97)*, pp. 214–223. Erasmus University Press, Rotterdam, The Netherlands, 1997.

Herman T. Tavani. Data Mining, Personal Privacy, and Public Policy. *Proceedings of the Conference on Computer Ethics: Philosophical Enquiry* (*CEPE '98*), PP. 113–120. The University of London Press, London, UK, 1998.