

Andrew Charlesworth
Director, Centre for IT & Law

Legal and ethical issues in data mining



Introduction

- Why are we interested in privacy?
- Larson's Laws of Data Dynamics
- Fair Information Principles
- Ethical Issues in Information Technology
- Data Mining
- Privacy and Ethical issues

Larson's Laws of Data Dynamics

- **The First Law:** Data must seek and merge with complementary data.
- **The Second Law:** Data always will be used for purposes other than originally intended.
- **The Third Law:** Data collected about individuals will be used to cause harm to one or more members of the group who provided the information or about whom it was collected, be it minor or major.
- **The Fourth Law:** Confidential information is confidential only until someone decides it's not.

Larson, Erik. *The Naked Consumer* (1992)

First Law

- As an individual you 'leak' data about yourself constantly - technology can turn that leak from a steady drip to a torrent
- Many organizations collect parts of that datastream to compile complex profiles: retailers - Sears, credit bureaus - Equifax, insurance companies - Medical Information Bureau, government.
- To those organizations and third parties who use their datasets, your collected personal data, or 'data shadow', is often effectively the real 'you'.
- But your 'data shadow' may not accurately reflect 'you'...

“The traffic in human information now is immense. There is almost nothing the commercial and governmental world is not anxious to find out about us as individuals.”

Privacy Commissioner of Canada, 1996

“Macleans was able to purchase the Privacy Commissioner's phone logs online from a U.S. data broker, no questions asked - detailed lists of the phone calls made from her Montreal home, Eastern Townships' chalet, and to and from her government-issued BlackBerry cellphone.

Macleans Nov 21, 2005

Second Law

- Collections of personal data are ‘valuable’ in all kinds of ways
- Consider, for example, a national DNA database:
 - the ultimate ID card
 - useful for law enforcement (scene of crime)
 - useful for law enforcement (prevention)
 - helpful for identifying public health trends
 - potentially valuable to insurance companies
- The more aggregated data there is in a collection, the broader the scope of potential uses, and the greater the pressure to broaden access to it.

- 2003 - Oyster card introduced on London, UK public transport network
 - 2004 - 7 requests by police for Oyster card information
 - 2006 - 1900 requests by police for Oyster card information
-
- 2003 - Congestion charging introduced to London, UK
 - Congestion charges are only in force at peak times, but the camera system runs 24 hours a day
 - 2007 - Police given live access to congestion charge cameras.



Third Law

- Secondary uses of personal data can cause harm, due to deliberate or inadvertent misuse.
- Consider the ways that apparently benign information about you could be used:
 - Politics/Religion - book purchases, library borrowing, memberships, blogs
 - Medical - medical records, prescriptions, e-mail lists
 - Other - video rentals, consumer surveys, loyalty cards, government records.
- Controlling access to, and use of, your personal data can be difficult.



- Dutch local administrative records pre-WW2 contained reference to the religion of local citizens.

- Vehicle Licensing Agencies are sources of much useful information



- As are pre-natal classes at hospitals and clinics, notably in the US

Fourth Law

- And this assumes that those who collect your personal data have defending your interests in personal data privacy at heart
- If it comes to a choice between breaking a promise to you or paying off debtors, or boosting profits, what are personal data collectors going to do?
- In the absence of rules forbidding them to trade in your personal data, then the odds on them maintaining your privacy don't look good.
- And then there's the issue of government requests for personal data...



- Toysmart.com filed for bankruptcy in June 2001. Its privacy policy said it would not share its customers' data with 3rd parties, but it tried to sell the data in bankruptcy.
- In 2004, FTC alleges Gateway Learning rents personal information provided by consumers – names, addresses, phone numbers, ages and gender of children – to marketers despite its privacy policy
- In 2005, FTC alleges Cart Manager Intl. rented personal information about merchants' customers to marketers, knowing that such disclosure contradicted merchant privacy policies.



- In 2002 US TSA requested that JetBlue turn over 5,000,000 passenger name records to the Department of Defense for various data mining/analysis. This violated JetBlue's privacy policy.
- From 2001 onwards, the US govt. required the *Society for Worldwide Interbank Financial Telecommunication* to provide info about international financial transactions, breaching several other nations' data privacy laws.
- US EFF alleges that AT&T has handed over customer phone records to the US National Security Agency. In 2006 AT&T responds by changing its terms and conditions.

Fair Information Principles

- Be accountable
- Identify the purpose of data collection
- Obtain consent
- Limit collection
- Limit use, disclosure and retention
- Be accurate
- Use appropriate safeguards
- Be open
- Give individuals access
- Provide recourse

Reasons

- **Purpose**
Clear reasons for collection of personal data reduce unintended secondary uses
- **Fairness**
Even where it is necessary to process personal data this should be done fairly
- **Transparency**
Individuals should have the right to know their data is being processed, by whom and for what purpose.

Laws and Rules

- The Fair Information Principles underlie most international agreements, national laws and codes of practice concerning personal data privacy
 - OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data 1980
 - HK Personal Data (Privacy) Ordinance 1995
 - EU Data Protection Directive 1995
 - UK Data Protection Act 1998
 - EU/US Safe Harbor Agreement 2000
 - CA Personal Information Protection and Electronic Documents Act (PIPEDA) 2000

Ethical Issues

- In theory, when designing new business models, administrative systems and IT tools, consideration of the FIPs should be part of the process.
- In practice, managers fail to understand privacy issues and technologists fail to incorporate privacy into their designs.
- Often this is simply an oversight (sometimes costly)
 - Human Resources Development Canada's (HRDC) Longitudinal Labour Force File (cancelled in 2000)
- However, in other cases the matter is less clear cut...

What's in a fingerprint?

- 2002 - UK school uses fingerprint technology instead of library cards (w/o parents' consent)
- 2005 - fingerprint technology for school meals
 - “The scanner records what the pupils order and parents can get a full report of this every month.”
- 2006 - fingerprint technology for attendance registration
 - "All data is retained in the school as part of our current database and will not be shared with any third party"
- What are the benefits and risks?
 - "I'm just disappointed our parents wouldn't let us be on the forefront of this technology"

Data Mining

- Extracting useful information from large data sets or databases
- Extraction of implicit, previously unknown, and potentially useful information from data
 - searching a database (or a set of coupled databases) for items with a specific combination of characteristics that does not correspond to a standard query
 - a search for (frequently occurring or otherwise significant) combinations of characteristics in a database or collection of databases - descriptive data mining
 - search for patterns to be used with the aim of predicting certain characteristics - predictive data mining

(Birrer 2005)

Parameters

- association
 - patterns where one event is connected to another event, e.g. purchasing a pen and purchasing paper
- sequence or path analysis
 - patterns where one event leads to another event, e.g. the birth of a child and purchasing nappies
- classification
 - identification of new patterns, e.g. coincidences between duct tape purchases & plastic sheeting purchases
- clustering
 - finding and visually documenting groups of previously unknown facts, e.g. geographic location/brand preferences
- forecasting
 - discovering patterns from which one can make reasonable predictions regarding future activities e.g. people who join an athletic club may take exercise classes.

(Seifert 2004)

Issues

- Data mining may reveal patterns and relationships, but does not tell the user the value or significance of these patterns
- Validity of the patterns discovered is dependent on how they compare to 'real world' circumstances
- Data mining can identify connections between behaviours and/or variables, but does not necessarily identify a causal relationship
- GIGO – inaccurate data will cause inaccurate results, the consequences of this will depend on the importance of the result to any planned cause of action.

Large data-mining projects

- US Total or Terrorism Information Awareness (TIA) program (now ended)
 - data mining of 'transaction space'
 - transaction space = financial, educational, travel, medical, veterinary, country entry, place/event entry, transport, housing, critical resources & communications.
 - looked for "connections between transactions" (passports, visas, work permits, driver's licenses, credit card, airline tickets, rental cars, gun purchases, chemical purchases) and "events" (arrest or "suspicious activities").
 - TIA's Wargaming the Asymmetric Environment (WAE)
 - TIA's Scalable Social Network Analysis (SSNA)
- Now CAPPS II & MATRIX

Large data-mining projects

- Icelandic deCODE project
 - Genetics database consisting of info. based on DNA samples from 70,000 volunteers linked with and cross-referenced to medical/healthcare records, genealogical records, and genetic information in the 3 separate databases of the Icelandic Healthcare Database (IHD).
 - Icelandic parliament's granted deCODE exclusive rights (for 12 years) to the information in the nation's health-records database - Act on a Health Sector Database 1998
 - Computerized data-mining techniques are used to find correlated genes and gene variations, simply by comparing databases of genetic samples and disease records.

Privacy Issues

- Terrorism Information Awareness (TIA)
 - Lack of legal protection against unlawful covert surveillance (avoids US Fourth Amendment safeguards)
 - Scale of use invisible to the general public unlike traditional ‘search and seizure’.
 - Inaccuracies and false positives
 - Mission creep – what this could also be used for...
 - Security of large scale data collections (internal and external)
 - Lack of clarity/understanding of the risks of data-mining
 - ‘Chilling effect’ of this type of data mining

Privacy Issues

- deCode project
 - Genetic information is highly sensitive/personal
 - Can be used for wide range of purposes e.g. insurance
 - Can target not just individuals, but those related to them
 - Can be used to ‘socially sort’ individuals
 - Can be used to make aggregate decisions which while not based on personal data (and thus perhaps caught by data privacy laws) will have an impact on individuals – DP laws tend to be concerned with privacy rights based on info about the individual not about a group
 - Data mined may be used as if it is personal data even though it’s not - Non-distributional group profiles based on probabilities & averages.