DIMACS/PORTIA Workshop on Privacy Preserving Data Mining

Data Mining & Information Privacy:

New Problems and the Search for Solutions

March 15th, 2004

Tal Zarsky

The Information Society Project,

Yale Law School

Introduction:

- Various privacy problems addressed in the public debate and technological discourse
- I will address what problems I see as critical
- Thereafter move to address solutions
- Examine which forms of privacy policy are adequate

Introduction

The Scope of my project:

- Limited to the “commercial realm” – mostly with regard to databases commercial entities have already been obtained

  [leaving aside government’s analysis of data to track criminal and terrorist activity]

- Focus on the privacy implications in the Internet setting

Introduction – Why the Internet?

- Collection:

  Omnipresent, Quantity leap, Quality leap
- Analysis:

  Digital environment, easy to “warehouse”
- Use: Narrowcasting, tailored content and the “feedback loop”

Introduction – Why the Internet?

Bringing it together: Amazon and the recommendation system

Bringing it together (2): AOL and the “walled garden”

Bringing it together (3): pen registers and wiretapping – the shift from the phone to the Internet – and from *Smith vs. Maryland* to the USA PATRIOT Act

In *conclusion*: The Internet is a very interesting test case and an opportunity to learn about policy implications in a wider setting as well.

Identifying the problems:

Form of analysis:

Addressing concerns in legal and social literature, and examining the implications of data mining applications on these issues.

Why does this matter?

Identifying the problems

The significance of understanding data mining:

- Generates confusion and is often used in the wrong context
- When understanding the *problems* data mining tools generate – we can construct tools that mitigate these concerns

Identifying the problems – *data mining applications:*

Key elements of data mining applications in the privacy context:

- Powerful tools of data analysis – with the ability to carry through descriptive and predictive tasks
- Non-hypothetically driven – less human decision-making and involvement
- It is very difficult to know what will be the final results of each analysis

Identifying the problems

Privacy concerns:

- Privacy is a “tricky” concept
- Identify three “mega” problems stemming from the collection of personal data:

  (1) Fear the data will be passed on to government (*will not address* – yet is a serious “fear” and aspect in any information privacy discussion)

  (2) Fear of the collection of personal data *per se* (collection on its own is bad enough)

  (3) Fear of the specific detriments stemming from the use of personal data (the “so what?” approach)

Identifying the problems – *Fear of Collection per se*

Specific concerns:

- Loss of control over data, self-monitoring, conformity, inability to form intimacy, loss of autonomy

Overall response – *social adaptation*

The role of Data Mining

Identyfing the problems: *Metaphors we live by*

The powerful metaphors (and the problems they cause):

“1984”

Kafka (“The Trial”, “The Castle”)

“Brave New World”

Bentham’s “Panopticon”

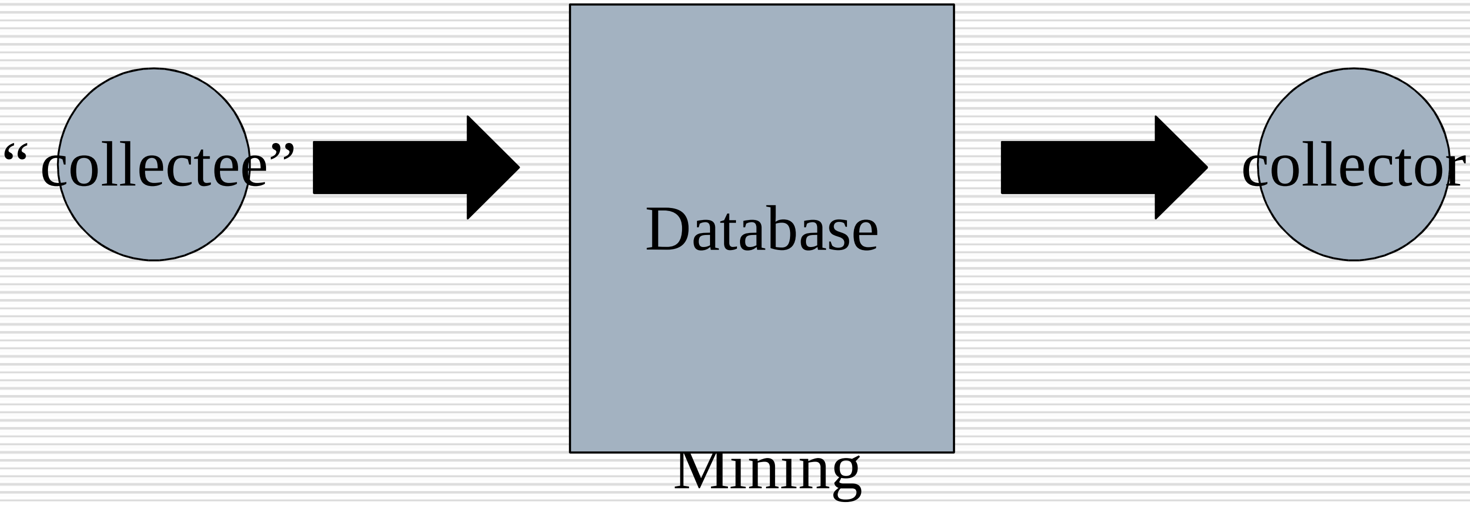
Common responses to “Privacy claims”

Privacy creates:

- Social costs: reputation, search expenses (waste)
- Security costs (inability to track terrorists, criminals, diseases)
- First Amendment claims (limitations on the transfer of personal information are a limitation of speech) – *U.S. West*

Leading thinkers: Posner, Etzioni, Cate

Identifying the problems: “The Tragedy of Errors”



1. Errors in the data
2. Errors in the process
3. (a) false positive
   (b) false negative
5. Human vs. Machine

“Tragedy of errors”

Errors in the Data:

- History: stems from “credit reporting” concerns
- Solution – access and correction (companies do not really object)
- Data Mining? Can mitigate concerns

“Tragedy of errors”

Errors in the process:

Drawing inferences leads to mistakes

Ms. Gray has received notice indicating that she would be charged a high premium for insurance. The facts accumulated by the insurance company with regard to Ms. Gray are all true: She subscribes to Scuba Magazine, visits Internet sites discussing bungi jumping, and travels each year to the Himalayas. Given these facts, the insurance firm concluded that Ms. Gray is a “risk-taker” and priced her policy accordingly. However, this conclusion is far from accurate, as Ms. Gray’s idea of risk-taking is buying blue chip stocks and boarding the subway after 6 p.m. She is currently writing an article about the dangers of extreme sports, and travels to Tibet to visit her son.

False positives & False negatives – different implications in different settings (*for example:* terrorism – false negative – devastating results)

Great deal of uncertainty – from neo- Luddite to healthy skepticism

Can data mining help or make things worse *(key issue to be examined!)*?

The “Human Touch”:

Is there specific importance in human participation in a decision making process?

Humans will identify instances where rules should be broken

Humans have biases. Data mining might help mitigate these concerns.

Back to the metaphors – *2001* (and now *the Matrix*)

Identifying the problems

- Abuse
- Discrimination:

  (1) In general

  (2) Problematic Factors

  (3) Based on prior patterns of behavior
- Autonomy and Manipulation

Identifying the problems “Abuse”

“*…a Los Angeles Man, Robert Rivera, says that after he sued Vons markets when he fell in the store and injured his leg the store looked up his record, discovered that he likes to buy a lot of liquor, and said it would use the information to defend itself in the lawsuit. The implication was that Rivera may have been impaired when he fell*.” Privacy Journal Mar. 1999, at 5.

Identifying the problems: Abuse

Fears in general:

Disclosure of facts, blackmail, embarrassment

Role of data mining – minimal (yet privacy preserving data mining tools might allow circumvention of these concerns)

Response in the “tort” structure:

“The Tort of Private Facts” – notoriously hard to establish

“Appropriation” – usually limited to commercial gains from name and face.

Identifying the problems *Discrimination*

Discrimination:

Treating consumers and users *differently* on the basis of their personal data

Different connotation in the legal and economics context

Discrimination is important:

* limits cross subsidy between consumers

* Additional consumers can enter market

* Firm surplus may lead to consumer surplus

Discrimination

Discrimination on the basis of “problematic factors”

- Law’s usually only concern government activities (in addition to some laws that concern private actors such as landlords – and “redlining”)
- This form of discrimination may prove to be a successful business strategy – and may not be motivated by animosity (Ayres) – indications of high transactional, searching and information costs
- The role of data mining: positive (limited bias in collection and analysis) or negative (lead to discrimination *de facto*)?
- Accepted forms of solutions: Collection of “sensitive data” is restricted

Discrimination

Using personal data to “efficiently” price products and services- on the basis of the users previous behaviors and preference

The role of data mining – extremely effective in identifying trends and predicting results

The problems: (1) Imbalance between the parties’ knowledge and risk when entering a transaction (2) lack of transparency

The privacy “challenge” – constructing tools that strike the balance between efficient business practices and unfair ones.

Autonomy:

- Difficult and problematic concept
- “insight” into the users preferences allows content providers to effectively *manipulate* them
- Realistic in view of applications such as the “Daily Me”
- Data Mining can play substantial role in view of ability to predict trends of behavior
- Again concerns go to the lack of *balancing* and *transparency*

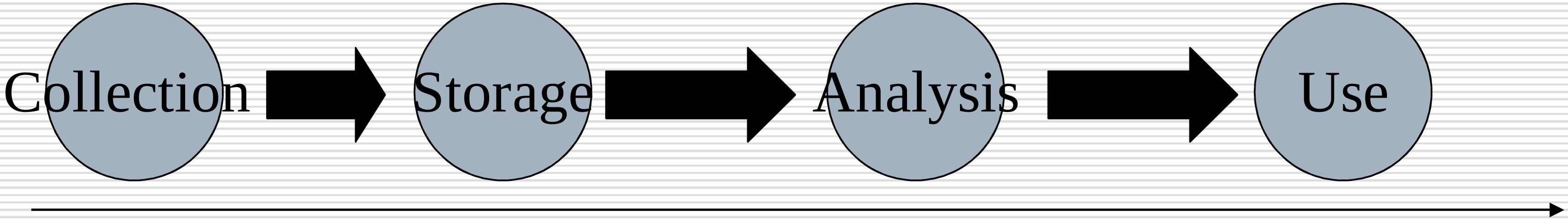
Overview of solutions

- "The Right of Privacy" (1890)
- Torts – the Four Privacy Torts (Prosser, 1960): Intrusion, Disclosure of Private Facts, False Light, Appropriation – garden variety of rights
- The EU Directive – and overall perspective (understanding *secondary sale & secondary Use*; *Opt In* vs. *Opt Out*)
- The Fair Information Practices – Notice, Access, Choice, Security and Enforcement
- The U.S. Patchwork –
  - Protected realms - Health (HIPPA)
  - Protected Subjects - Children (COPPA)
  - Protected forms of Data ("Sensitive Data")

Why Torts (usually) fail – and the realm of today's data collection

  - Example: DoubleClick and "cookies"
- The contractual and property perspective (*for example:* default and mandatory rules)
  - The technological solution (P3P, Lessig)
- The shortcoming – and the implications of data mining
  - Market failures (high information and transactional costs) – *people are happy to sell their privacy for very very cheap!*
  - Negative externalities (inferences from one group to another, and from group to individual
  - Loss of Benefits (loss of subsidy to start ups, loss of data derived from analysis)

The Flow of Personal Data



Collection

Storage

Analysis

Use

Solutions – now what?

Understanding the benefits of data analysis – and concentrating on applications on the “implementation” end of the data flow

Examining the role of transparency and pseudonymity

Embedding values in accepted protocols for the analysis of personal data