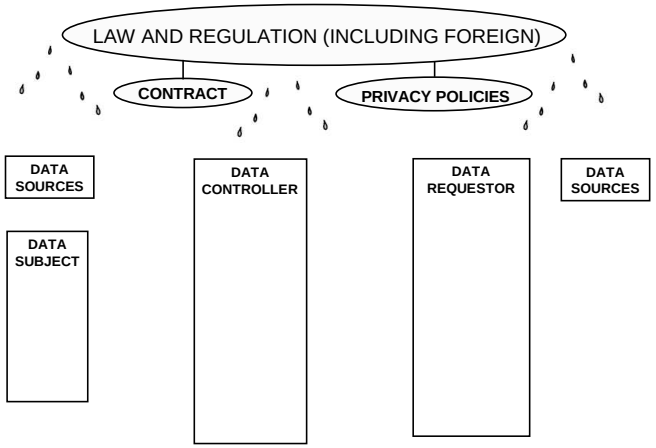
Mathematics and the Privacy Laws

Michael I. Shamos, Ph.D., J.D.
Co-Director, Institute for eCommerce
Carnegie Mellon University

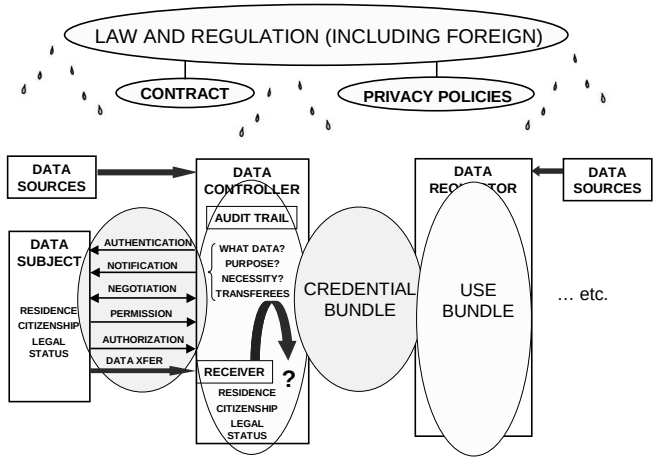
Institute for eCommerce

PRIVACY: IT'S NOT JUST A GOOD IDEA ….

International Privacy Model



International Privacy Model



Data Privacy Research

- Privacy of databases
  - Privacy of de-identification transformations
- Privacy specification
  - Machine-readable privacy constraints
  - Relation to law
  - Semantics of data
- Privacy administration
  - How can privacy constraints be enforced by machines?

Privacy Act of 1974

- Applies to federal agencies
- "No agency shall disclose any record … to any person, or to another agency, except … with the prior written consent of, the individual to whom the record pertains, unless disclosure of the record would be --
  - … used solely as a statistical research or reporting record, and the record is to be transferred in a form that is not individually identifiable" (not a defined term)
- Restriction on "matching programs"
  - any computerized comparison of -- (i) two or more automated systems of records … [certain exceptions]

5 USC §552(b)

Gramm-Leach-Bliley 15 USC §5801ff

- Except as … authorized …, you may not, directly or through any affiliate, disclose any nonpublic personal information about a consumer to a nonaffiliated third party unless:
  (i) You have provided to the consumer an initial notice as required;
  (ii) You have provided to the consumer an opt out notice as required in Sec. 313.7
  (iii) You have given the consumer a reasonable opportunity, before you disclose the information to the nonaffiliated third party, to opt out of the disclosure; and (iv) The consumer does not opt out.

  16 CFR §313.8
- Applies to "financial institutions," a very broad category

What Gramm-Leach-Bliley Protects

- "Nonpublic personal information" means:
  (i) Personally identifiable financial information; and
  (ii) Any list, description, or other grouping of consumers (and publicly available information pertaining to them) that is derived using any personally identifiable financial information that is not publicly available.

  TRANSITIVE CLOSURE
- "Personally identifiable financial information" means any information:
  (i) A consumer provides to you to obtain a financial product or service;
  (ii) About a consumer resulting from any transaction involving a financial product or service between you and a consumer; or
  (iii) You otherwise obtain about a consumer in connection with providing a financial product or service to that consumer.

  EXTERNAL DATA

  16 CFR §313.3

No Sharing of Account Numbers

- You must not, directly or through an affiliate, disclose, other than to a consumer reporting agency, an account number or similar form of access number or access code for a consumer's credit card account, deposit account, or transaction account to any nonaffiliated third party for use in telemarketing, direct mail marketing, or other marketing through electronic mail to the consumer.

16 CFR §313.14

USA Patriot Act

- Reports relating to coins and currency received in nonfinancial trade or business
- "Any person -- (1) who is engaged in a trade or business; and … receives more than $10,000 in coins or currency in 1 transaction (or 2 or more related transactions), shall file a report … with the Financial Crimes Enforcement Network
  [including] the name and address … of the person from whom the coins or currency was received"

  31 USC §5331
- Immunity for reporting "suspicious activities"
- Immunity for including "suspicions of illegal activities in written employment references"

What HIPAA Provides

- A covered entity may not use or disclose protected health information, except as permitted or required …
  - pursuant to … a consent … to carry out treatment, payment, or health care operations
  - pursuant to … an authorization
  - pursuant to … an agreement (opt-in)
  - [other provisions]

  45 CFR §164.502
- Health information that meets … specifications for de-identification … is considered not to be individually identifiable health information

  45 CFR §164.502(d)

Compliance deadline April 14, 2003

What HIPAA Protects

- "Individually identifiable health information" is information that is a subset of health information, including demographic information collected from an individual, and: …
  - relates to … physical or mental health or condition of an individual;
    … provision of health care to an individual; or
    … payment for … health care to an individual; and
  - identifies the individual; or
  - with respect to which there is a reasonable basis to believe the information can be used to identify the individual

  45 CFR §164.501

De-Identification

- **A covered entity may determine that health information is <u>not individually identifiable</u> only if: … the following identifiers of the individual or of relatives, employers, or household members of the individual are removed:**
- **Names;**
- **All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, …, except for the initial three digits of a zip code if …**
- **All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89…**
- **Telephone numbers; Fax numbers; email addresses; URLs; IP addresses**
- **Social security numbers; Medical record numbers; Health plan beneficiary numbers; Account numbers;**
- **Certificate/license numbers; vehicle identifiers, serial numbers, plate numbers;**
- **Device identifiers and serial numbers;**
- **Biometric identifiers, including finger and voice prints;**
- **Full face photographic images and any comparable images; and**
- **Any other unique identifying number, characteristic, or code; and**
- **The covered entity does not have <u>actual knowledge</u> that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.**

<u>45 CFR §164.514</u>

Mathematics of De-Identification

- A function f is *k-private* iff no finite set of argument-value pairs $\{(x_1,f(x_1)), \ldots, (x_n,f(x_n))\}$ suffices to compute f at any other point $x \notin \{x_1, \ldots, x_n\}$.
- Example: a polynomial of degree exactly k is k-private but not $(k+1)$-private
- Wait. This depends on <u>knowing</u> that f is a polynomial of degree k. That is <u>metadata</u> about f (semantics).
- What kinds of metadata can we have about functions and what can be inferred from metadata?
- A function f is *totally private* iff it is k-private for all $k>0$
- Example: a general power series $p(x) = \sum_{i=0}^{\infty} a_i x^i$ is totally private (without other metadata about $p(x)$)
- In general, k-privacy and total privacy are undecidable

De-Identification Through Encryption

- Encryption is reporting the value $g(f(x))$ instead of $f(x)$
- g is 1-1 so it has an inverse (though the inverse may not be known)
- g should be totally private; otherwise it can be compromised
- g is a *derangement function* if it has no fixed point, i.e.,

$$\nexists x \ s.t. \ f(x) = x$$

- Derangement alone is not enough
  - Need to avoid short periods: $g(g(x)) = x$
  - Also want g to be computationally one-way. $g(x) = x+1$ is no good
- Whether g has a fixed point is undecidable. Generalized SHA?
- Question: when is $g(f(x))$ totally private?
- There exists a totally private derangement function g and a totally private f such that $g(f(x))$ is determined everywhere

Which Function Properties Are Decidable?

- Suppose $f(x)$ is finitely presented. Which of its properties are decidable?
- Example, let $f(x)$ be a polynomial
- Is it decidable whether $\sum_{k=1}^{\infty} \frac{1}{f(k)}$ is rational?
- Examples: $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$ $\sum_{k=1}^{\infty} \frac{1}{k^2+2k} = \frac{3}{4}$ $\sum_{k=1}^{\infty} \frac{1}{k^2+1} = \frac{\pi \coth \pi - 1}{2}$

$\sum_{k=1}^{\infty} \frac{1}{k^3} = \zeta(3) = \textit{irrational}$ $\sum_{k=1}^{\infty} \frac{1}{k^5} = \zeta(5) = \textit{rationality unknown}$

Generalizations

- Distortion (a type of generalization): suppose instead of $f(x)$ we report a random number from a distribution D_f
- Example: choose a value y uniformly in the interval $[f(x)-a, f(x)+a]$
- Suppose we consider $f(x)$ identified if $|y - f(x)| \leq \varepsilon$
- Properties of f and D may enable us to bound $f(z)$ (a different value) definitively
- Properties of f and D may enable us to bound $f(z)$ probabilistically

Major Ideas

- Privacy involves
  - data semantics
  - expression languages for authorizations, permissions, purposes, status of parties, laws, contracts, policies
  - audit mechanisms
- Privacy laws are ambiguous and inconsistent, but must be obeyed
- Privacy problems lead to interesting mathematics

References

- Fischer-Hübner, S. & Ott, A. From a Formal Privacy Model to its Implementation, *Proc. 21st National Information Systems Security Conf.*, Arlington, VA, October 5-8, 1998.
- Hughes, D. & Shmatikov, V. Information Hiding, Anonymity and Privacy: A Modular Approach, *J. Comp. Security*, to appear
- Sweeney, L. *k*-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 557-570.
- Sweeney, L. Achieving *k*-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 571-588.