# Sensitive Information in a Wired World

**Supported by the National Science Foundation under the ITR Program**

**JOAN FEIGENBAUM**

**http://www.cs.yale.edu/homes/jf**

# **PORTIA**: Privacy, Obligations, and Rights in Technologies of Information Assessment

Large-ITR project described in NSF proposal as a "five-year, multi-institutional, multi-disciplinary, multi-modal research project on end-to-end handling of sensitive information in a wired world"

# Ubiquity of Computers and Networks Heightens the Need to Distinguish

- Private information
  - Only the data subject has a right to it.
- Public information
  - Everyone has a right to it.
- Sensitive information
  - "Legitimate users" have a right to it.
  - It can harm data subjects, data owners, or data users if it is misused.

# Examples of Sensitive Information

- Copyright works
- Certain financial information
  - Graham-Leach-Bliley uses the term "nonpublic personal information."
- Health Information

  Question: Should some information now in "public records" be reclassified as "sensitive"?

# State of Technology

+ We have the ability (if not always the will) to prevent *improper access* to private information. Encryption is very helpful here.

- We have little or no ability to prevent *improper use* of sensitive information. Encryption is less helpful here.

# PORTIA Goals

- Produce a next generation of technology for handling sensitive information that is qualitatively better than the current generation's.
- Enable end-to-end handling of sensitive information over the course of its lifetime.
- Formulate an effective conceptual framework for policy making and philosophical inquiry into the rights and responsibilities of data subjects, data owners, and data users.

# Academic–CS Participants

Stanford

Dan Boneh

Hector Garcia-Molina

John Mitchell

Rajeev Motwani

Yale

Joan Feigenbaum

Ravi Kannan

Avi Silberschatz

Univ. of NM

Stephanie Forrest

("computational immunology")

Stevens

Rebecca Wright

NYU

Helen Nissenbaum

("value-sensitive design")

# Multidisciplinarity on Steroids

J. Balkin (Yale Law School)
G. Crabb (Secret Service)
C. Dwork (Microsoft)
S. Hawala (Census Bureau)
B. LaMacchia (Microsoft)
K. McCurley (IBM)
P. Miller (Yale Medical School)
J. Morris (CDT)
B. Pinkas (Hewlett Packard)
M. Rotenberg (EPIC)
A. Schäffer (NIH)
D. Schutzer (CitiGroup)

Note participation by the software industry, key user communities, advocacy organizations, and non-CS academics.

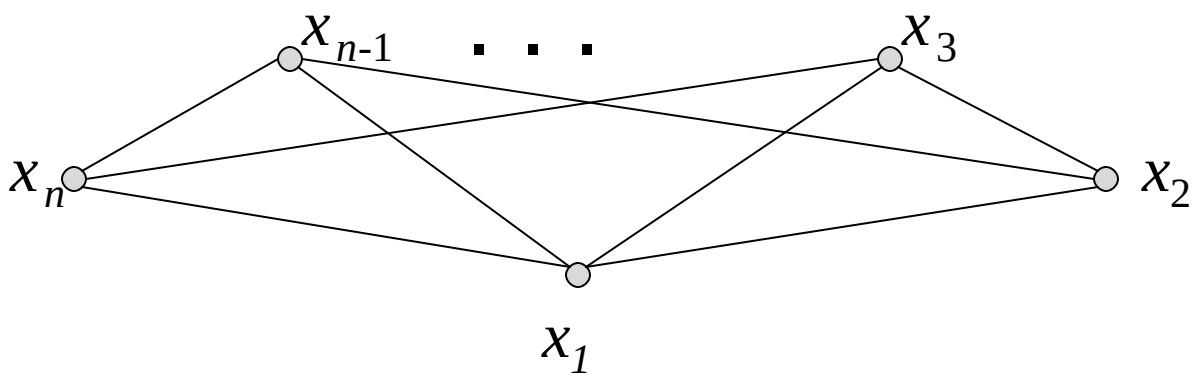
# Five Major Research Themes

- Privacy-preserving data mining and privacy-preserving surveillance
- Sensitive data in P2P systems
- Policy-enforcement tools for db systems
- Identity theft and identity privacy
- Contextual integrity

# Privacy-preserving Data Mining

- Is this an oxymoron?
- No! Cryptographic theory is extraordinarily powerful, almost paradoxically so.
- Computing exactly one relevant fact about a distributed data set while concealing everything else is exactly what cryptographic theory enables *in principle*. But not (yet!) in practice.

# Secure, Multiparty Function Evaluation




$$y = F(x_1, \ldots, x_n)$$

- Each *i* learns *y.*
- No *i* can learn anything about $x_j$ (except what he can infer from $x_i$ and $y$ ).
- Very general positive results. Not very efficient.

# PPDM Work by PORTIA-related Researchers

- Lindell and Pinkas: Efficient 2-party protocol for ID3 data mining on $x_1 \cup x_2$.
- Aggarwal, Mishra, and Pinkas: Efficient n-party protocol for order statistics of $x_1 \cup \cdots \cup x_n$.
- Freedman, Nissim, and Pinkas: Efficient 2-party protocol for $x_1 \cap x_2$.

# Some Areas in which Law and Technology Affect Each Other

- Internet access to “public records”
- Identification technology
- Unsolicited email and phone calls
- Digital copyright and DRM

# "Public Records" in the Internet Age

Depending on State and Federal law, "public records" can include:

- Birth, death, marriage, and divorce records
- Court documents and arrest warrants (including those of people who were acquitted)
- Property ownership and tax-compliance records
- Driver's license information
- Occupational certification

They are, by definition, "open to inspection by any person."

# How "Public" are They?

Traditionally: Many public records were "practically obscure."

- Stored at the local level on hard-to-search media, *e.g.*, paper, microfiche, or offline computer disks.
- Not often accurately and usefully indexed.

Now: More and more public records, especially Federal records, are being put on public web pages in standard, searchable formats.

# What are "Public Records" Used For?

In addition to straightforward, known uses (such as credential checks by employers and title searches by home buyers), they're used for:

- Commercial profiling and marketing
- Dossier compilation
- Identity theft and "pretexting"
- Private investigation
- Law enforcement

# Questions about Public Records in the Internet Age

- Will "reinventing oneself" and "social forgiveness" be things of the past?
- Should some Internet-accessible public records be only conditionally accessible?
- Should data subjects have more control?
- Should data collectors be legally obligated to correct mistakes?

# Identification Infrastructure Today I

- We are often asked to "present gov't-issued photo ID."
  - Airports
  - Buildings
  - Some high-value financial transactions
- Many gov't-issued photo IDs are easily forgeable.
  - Drivers' licenses
  - Passports
- We are often asked to provide personally identifying information (PII).
  - Social security number
  - Mother's maiden name
  - Date of birth
- Many people and organizations have access to this PII.

# Identification Infrastructure Today II

- Security of "foundation documents" (*e.g.*, birth certificates) is terrible.
- According to the US Department of Justice, the rate of identity theft is growing faster than that of any other crime in the United States.
- Existing technology could improve, if not perfect, ID security, *e.g.*:
  - Biometrics
  - Cryptographic authentication
- There is extensive research interest in improving this technology (and the *systems* that support it).

# Are Standard, Secure ID Systems Desirable?

+ Ordinary people could benefit from accurate, efficient identification, and identity thieves would have a harder time.
- Multi-purpose, electronic IDs facilitate tracking, linking, dossier compilation, and all of the other problems currently facilitated by Internet-accessible “public records.”
- Multi-purpose, standard “secure” IDs magnify the importance of errors in ID systems.

# Possible Approaches

- Build secure ID systems that *don't* facilitate linking and tracking.
  - Tracking a "targeted" person should require a court-ordered key.
  - Tracking someone for whom one doesn't have such a key should be provably infeasible.
  - There's already a plausible start on this in the security-theory literature.
- Organizations could "seize the high ground" by not retaining usage data for identification and authorization tokens (*a fortiori* not mining, selling, or linking it).
  - At least one ID start-up company is making this claim.
  - How can such a claim be proven?
  - Security theory does not address this question (yet!).

# What May We Use To Prevent Unwanted Phone Calls?

+ Technology
  - Answering machines
  - Caller ID
+ Money (together with technology)
  - “Privacy-guard service” from SNET
? Government
  - “Do-Not-Call” lists seem to be controversial.

# What May We Use To Prevent Unwanted Email?

+ Technology
  - Filters
  - CAPTCHAs
  - “Computational postage”
? Government
  + Yes, if the unwanted email is “trespass to chattel,” which requires that it “harm” the recipient’s computer system. (CyberPromotions)
  - No, if the email is merely “unwanted.” (Hamidi)

# Is a Network like a Country?

- Size, diversity, and universal connectivity imply risk. Get over it!
- Subnetworks ~ neighborhoods (J Yeh, CS457)
  - Some segregation happens naturally.
  - Gov't-sanctioned segregation is wrong.
- Alternative: Network nodes ~ homes (JF)
  - A man's computer is his castle.
  - Do I have to be rich or tech-savvy to deserve control over my own computer?

# Is there a Limit to the Upside of Network Effects?

Metcalf's Law: The value to a potential user of connecting to a network grows as the square of the number of users already connected.

Feigenbaum's Law: Metcalf's Law holds only until almost all potential users, including the scum of the earth, are connected. Then the value of the network drops to zero for almost everybody.

# Copyright: Dual Doomsday Scenarios

Today's Rights Holders and Distributors: Technical Protection Systems (TPSs) won't work. Copying, modification, and distribution will become uncontrollable.

Fair-Use Advocates: TPSs will work. Rights holders will have *more* control than they do in the analog world.

My Prediction: Both and neither! Copyright law, business models, TPSs, and users will evolve.

# Content-Distribution System Specification

- Part of the spec should be "enforce copyright law" (or at least "obey copyright law").
- In US Copyright Law
  + Owners are given (fairly) well defined rights.
  - Users are given "exceptions" to owners' rights.
- This is no way to specify a system!
- Need affirmative, direct specification of what users are allowed to do.

# What if Someone Builds a Good TPS?

- Lots of clever arguments in favor of
  - Users' rights to reverse engineer
  - Users' rights to circumvent
- These arguments are correct but insufficient
  - As system engineering (see "specification" slide).
  - As a philosophical position: If fair use is a part of the copyright bargain, then one should not have to hack around a TPS to make fair use.
  - As protection against ever-expanding rights of owners: What if someone builds a TPS that, for all practical purposes, can't be hacked?

# Content-Distribution System Engineering

- "Fair use analysis therefore requires a fact intensive, case-by-case approach." [Mulligan and Burstein 2002]
- This is no way to engineer a mass-market system!
- Need to be able to recognize the typical, vast majority of fair uses extremely efficiently and permit them.
- Note that, in the analog content-distribution world, the vast majority of fair uses *are* non-controversial.

# The Way Forward? I

- Rewrite copyright law so that it makes sense in today's (or any?) technological world.
  - + Preserve the *policy* ("Promote progress in science and the useful arts…").
  - - Change the technologically out-of-date *mechanisms (e.g.*, copy control).
- Sanity check: Create something that works as well for Internet-based content distribution as today's copyright law works for (physical) books.

# The Way Forward? II

- "The best TPS is a Great Business Model." [Lacy, Maher, and Snyder 1997]
- Use technology to *do what it does naturally*.
- An Internet content-distribution business should *benefit* from *uncontrolled copying and redistribution.*

# Core Technical Problem: The Unreasonable Effectiveness of Programmability

- Many machine-readable permissions systems
  - Rights-management languages
  - Privacy policies
  - Software licenses
- None is now technologically enforceable.
  - All software-only permissions systems can be circumvented.
  - Once data are transferred, control is lost.

# Will "Trusted Systems" Help?

- Hardware-based, cryptographic support for proofs that a data recipient's machine is running a particular software stack.
- Potential problems:
  - Technical: *Very* hard to build.
  - Business: Adoption hurdles.
  - Philosophy: Privacy, fair use, MS hatred, *etc*.
- Potential benefits:
  - Copyright enforcement? Maybe.
  - Privacy enforcement? *Much* harder!