Sensitive Information in a Wired World

CPSC 457/557, Fall 2013

Lecture 1, August 30, 2013

1:00-2:15 pm; AKW 400

http://zoo.cs.yale.edu/classes/cs457/fall13/

What is "Sensitive Information"?

Information that can harm data subjects, data owners, or data users if it is used improperly.

Note that not all sensitive information is "private" as that word is intuitively understood.

This course was inspired by the PORTIA project: http://crypto.stanford.edu/portia/

Course Requirements and Grading

- Reading assignments: The assigned reading will be discussed in class, and your participation in these discussions will be the basis for 10% of your course grade*.
- Oral presentation on a "sensitive-information" topic of your choice, worth 30% of your course grade; graduate students must also submit a paper on the same topic.
- 2 In-Class Exams (Oct. 15 and Dec. 5), each worth 30% of your course grade
- No final exam during exam week

* **(There's a way around this.)**

Instructor: Joan Feigenbaum
Office: AKW 512
Office Hours: Thur. 11:30 am - 12:30 pm
and by appointment
Phone: 203-432-6432
Assistant: Judi Paige
(**judi.paige@yale.edu**, 203-436-1267,
AKW 507a, 8:30 am – 4:30 pm M-F)

Note: Please don't expect emailed replies from Professor Feigenbaum, who suffers from RSI. Cc all email to Ms. Paige, who will reply (but not immediately).

If you are unsure about whether to take this course, peruse the PORTIA website and the course websites from 2011, 2006, and 2003.

Questions?

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

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

http://crypto.stanford.edu/portia/

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/Rutgers

Rebecca Wright

NYU

Helen Nissenbaum

("value-sensitive design")

Highly Multidisciplinary

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.

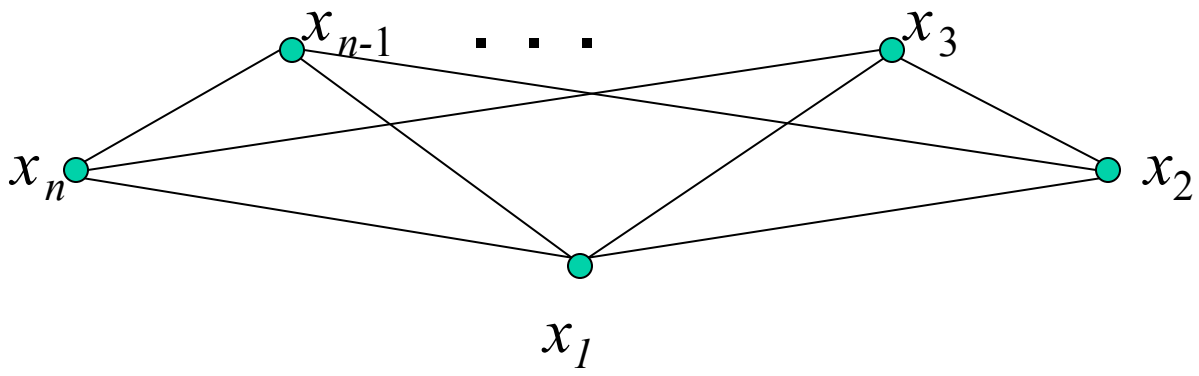
Five Major Research Themes

- 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.

Secure Computation of Surveys

Joan Feigenbaum (Yale), B. Pinkas (HP),

R. Ryger (Yale), and F. Saint-Jean (Yale)

http://www.cs.yale.edu/homes/jf/SMP2004.{pdf, ppt}

Surveys and other Naturally Centralized Multiparty Computations

- Consider
  - Sealed-bid auctions
  - Elections
  - Referenda
  - Surveys
- Each participant weighs the hoped-for payoffs against any revelation penalty ("loss of privacy") and is concerned that the computation be fault-free and honest.
- The implementor, in control of the central computation, must configure auxiliary payoffs and privacy assurances to encourage (honest) participation.

CRA Taulbee Survey: Computer Science Faculty Salaries

- Computer science departments in four tiers, 12 + 12 + 12 + all the rest
- Academic faculty in four ranks: full, associate, and assistant professors, and non-tenure-track teaching faculty
- Intention: Convey salary distribution statistics per tier-rank to the community at large without revealing department-specific information.

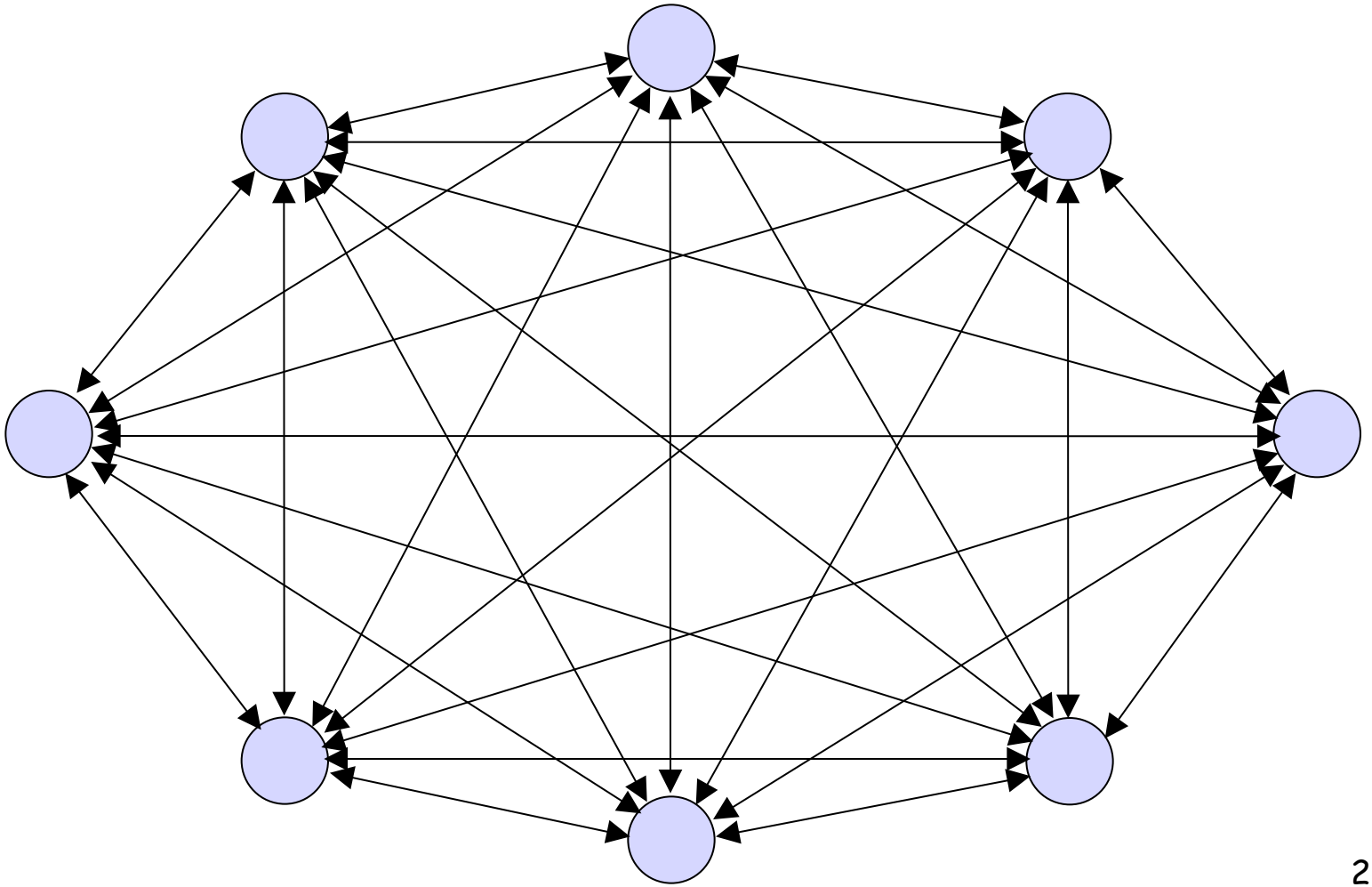
CRA Taulbee Survey: The Current Computation

- Inputs, per department and faculty rank:
  - Minimum
  - Maximum
  - Median
  - Mean
- Outputs, per tier and faculty rank:
  - Minimum, maximum, and mean of department minima
  - Minimum, maximum, and mean of department maxima
  - Median of department means (not weighted)
  - Mean (weighted mean of department means)

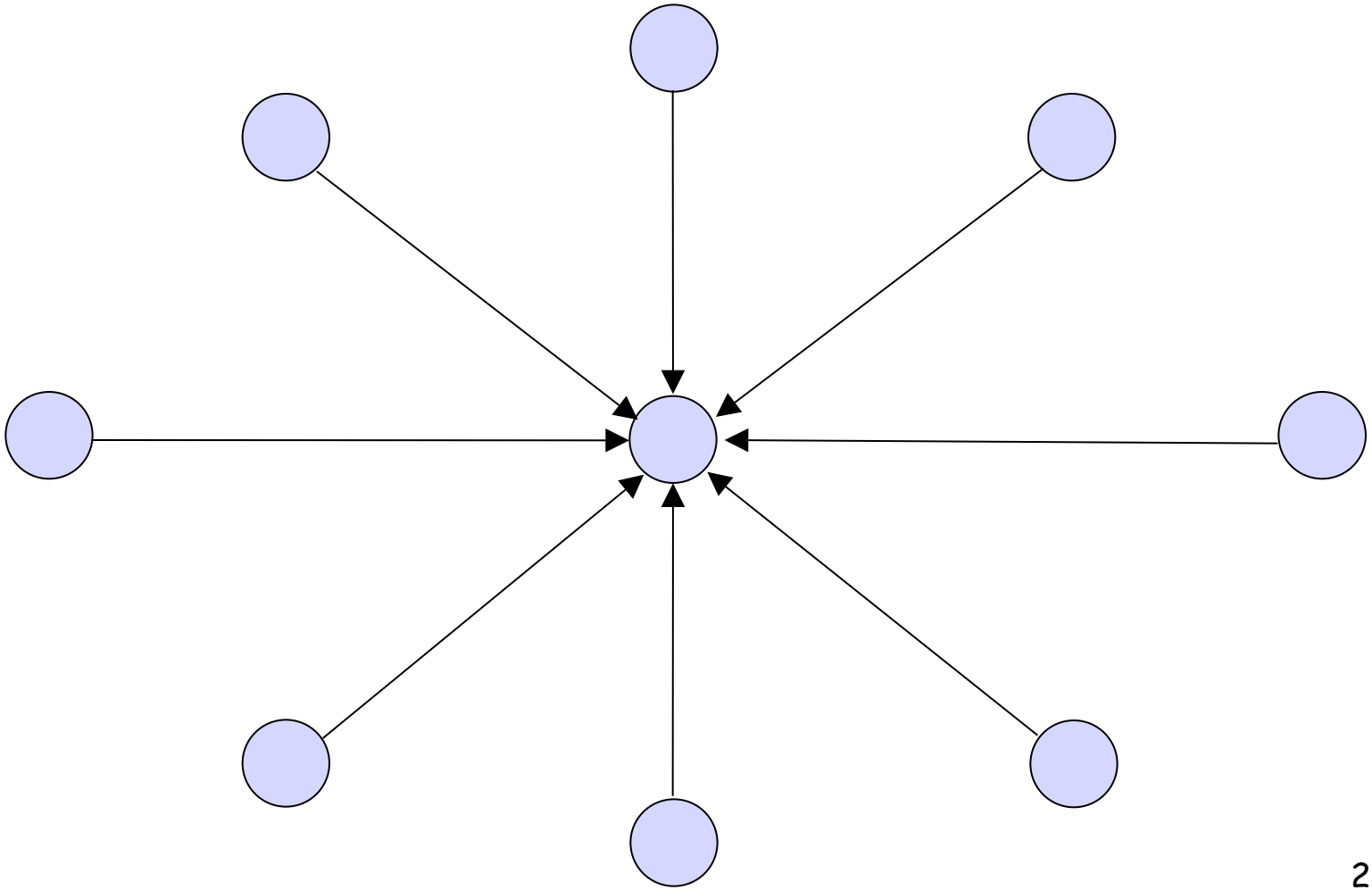
Taulbee Survey: The Problem

- CRA wishes to provide *fuller statistics* than the meager data currently collected can support.
- The current level of data collection *already compromises department-specific information*. Asking for submission of full faculty-salary information greatly raises the *threshold for trust* in CRA's intentions and its security competence.
- Detailed disclosure, even if anonymized, may be explicitly prohibited by the school.
- Hence, there is a danger of significant *non-participation* in the Taulbee Survey.

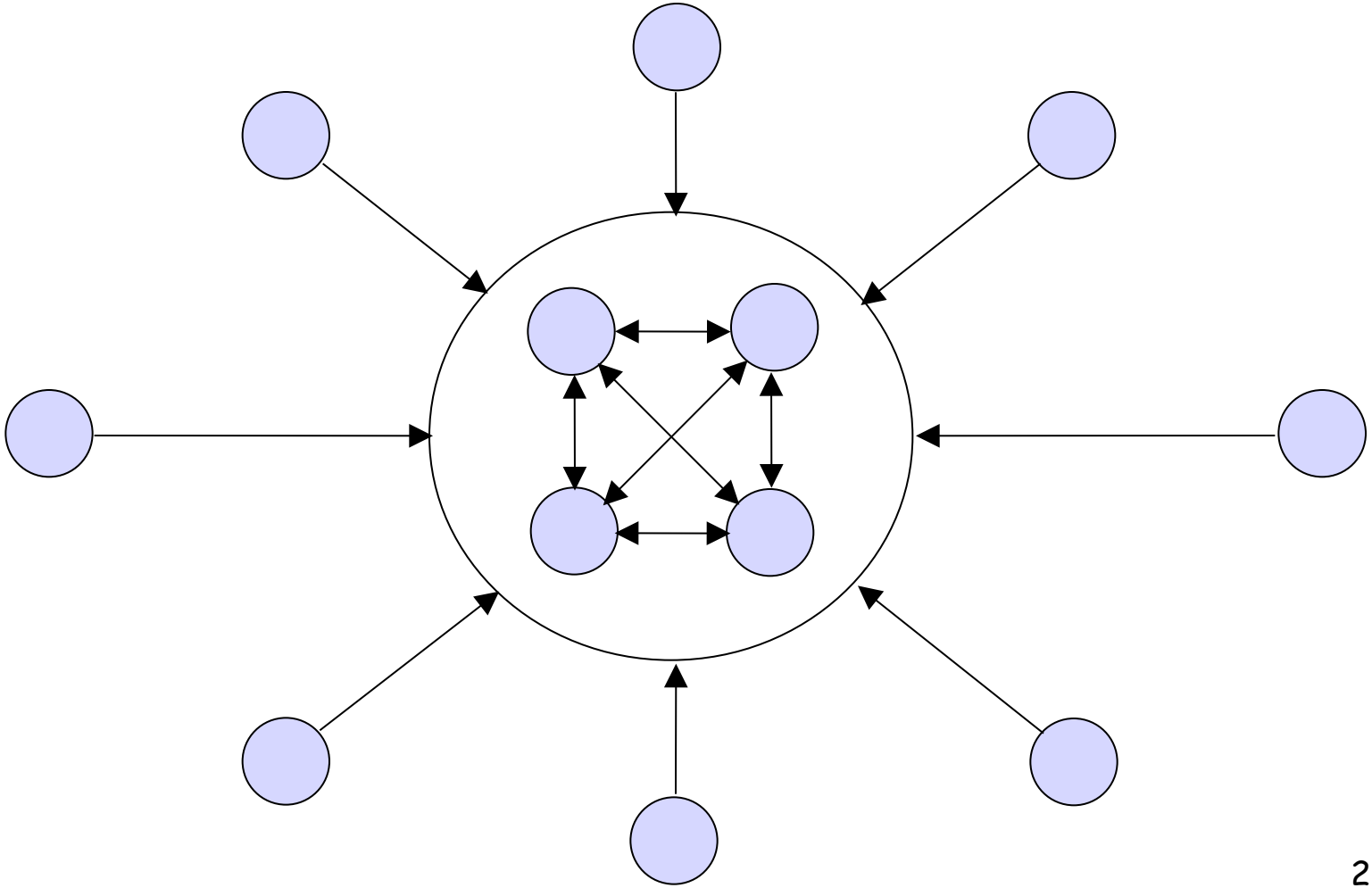
Communication Pattern: General SMFE Protocols



Communication Pattern: Surveys and Other Trusted-Party Computations



Communication Pattern: *M*-for-*N*-Party SMFE



Our Implementation: Input-Collection Phase

- Secure input collection:
  - Salary and rank *data entry* by department representative
  - Per rank, in JavaScript, *computation of XOR shares* of the individual salaries for the two ($M = 2$) computation servers
  - Per rank, *HTTPS transmission* of XOR shares to their respective computation servers
- Note that cleartext data *never leave the client machine.*

Our Implementation: Computation Phase

- Per tier and rank, *construction of a Boolean circuit* to
  - reconstruct inputs by XOR-ing their shares
  - sort the inputs in an odd-even sorting network
- Secure computation, per tier and rank:
  - *Fairplay* [Malkhi *et al.*, 2004] implementation of the *Yao 2-party SFE* protocol for the constructed circuit and the collected input shares
  - Output is a sorted list of all salaries in the tier-rank.
- Postprocessing, per tier and rank:
  - *arbitrary, statistical computation* on the sorted, cross-departmental salary list

The Heartbreak of Cryptography

- User-friendly, open-source, *free* implementation
- NO ADOPTION !@%$#
- CRA's reasons
  - ∗Need for data cleaning and multiyear comparisons
  - – Perhaps most member departments will trust us.
- Yale Provost's Office's reasons
  - ∗No *legal* basis for using this privacy-preserving protocol on data that we otherwise don't disclose
  - ∗Correctness and security claims are hard and expensive to assess, despite open-source implementation.
  - ∗All-or-none adoption by Ivy+ peer group.

PWS: A privacy application for Web search

Felipe Saint-Jean

joint work Aaron Johnson, Dan Boneh, and Joan Feigenbaum

ACM WPES 2007

Sensitivity of searches: an example

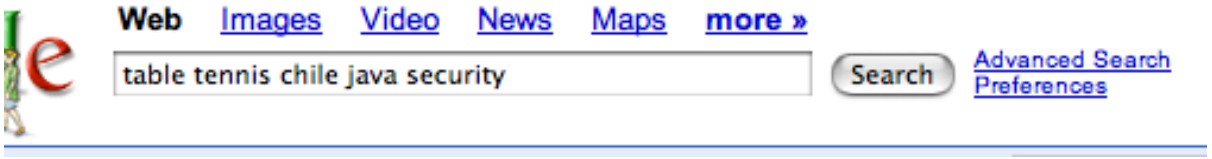
Search history

"Table Tennis Tournament New York"

"Java reflection"

"Chilean bakery new york"

"names buffer overflow"



Sensitivity of searches: an example

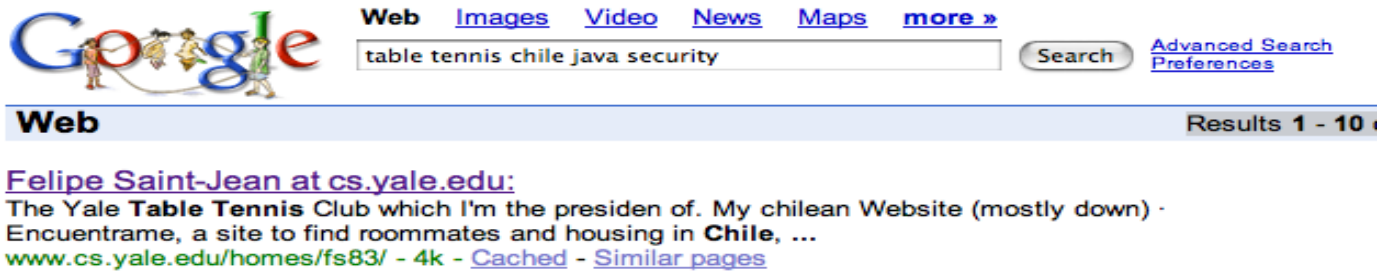
Search history

"Table Tennis Tournament New York"

"Java reflection"

"Chilean bakery new york"

"names buffer overflow"



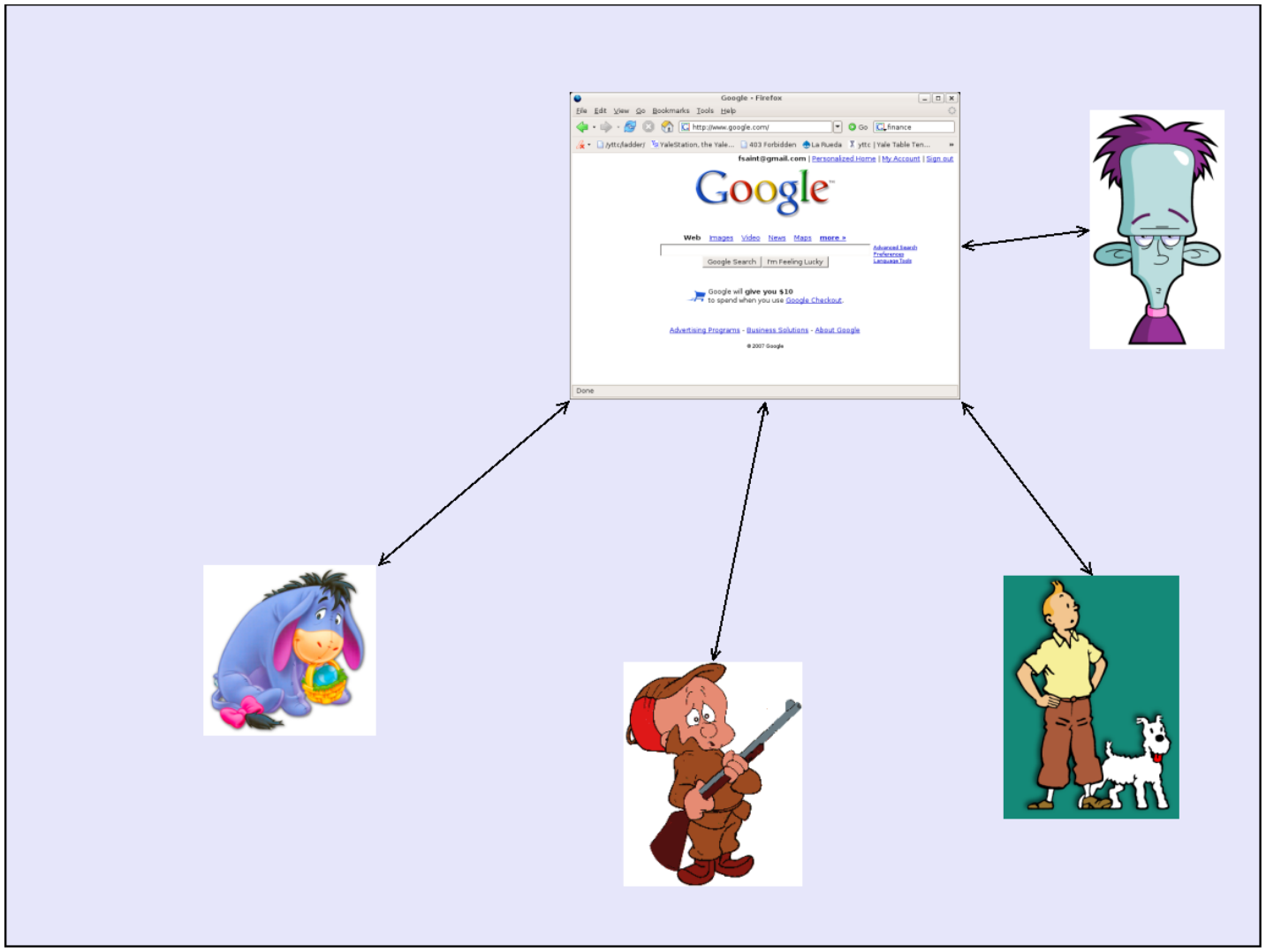
What information does the search engine collect?

- TCP/IP
  - IP address
  - Institution of ISP
  - OS
  - uptime
- HTTP Headers
  - Cookies
  - Operating system and OS version
  - Browser make and version
  - Encodings and language
- HTML
  - JavaScript collected information
  - Timing information
- Query terms and time
- Active components
  - …

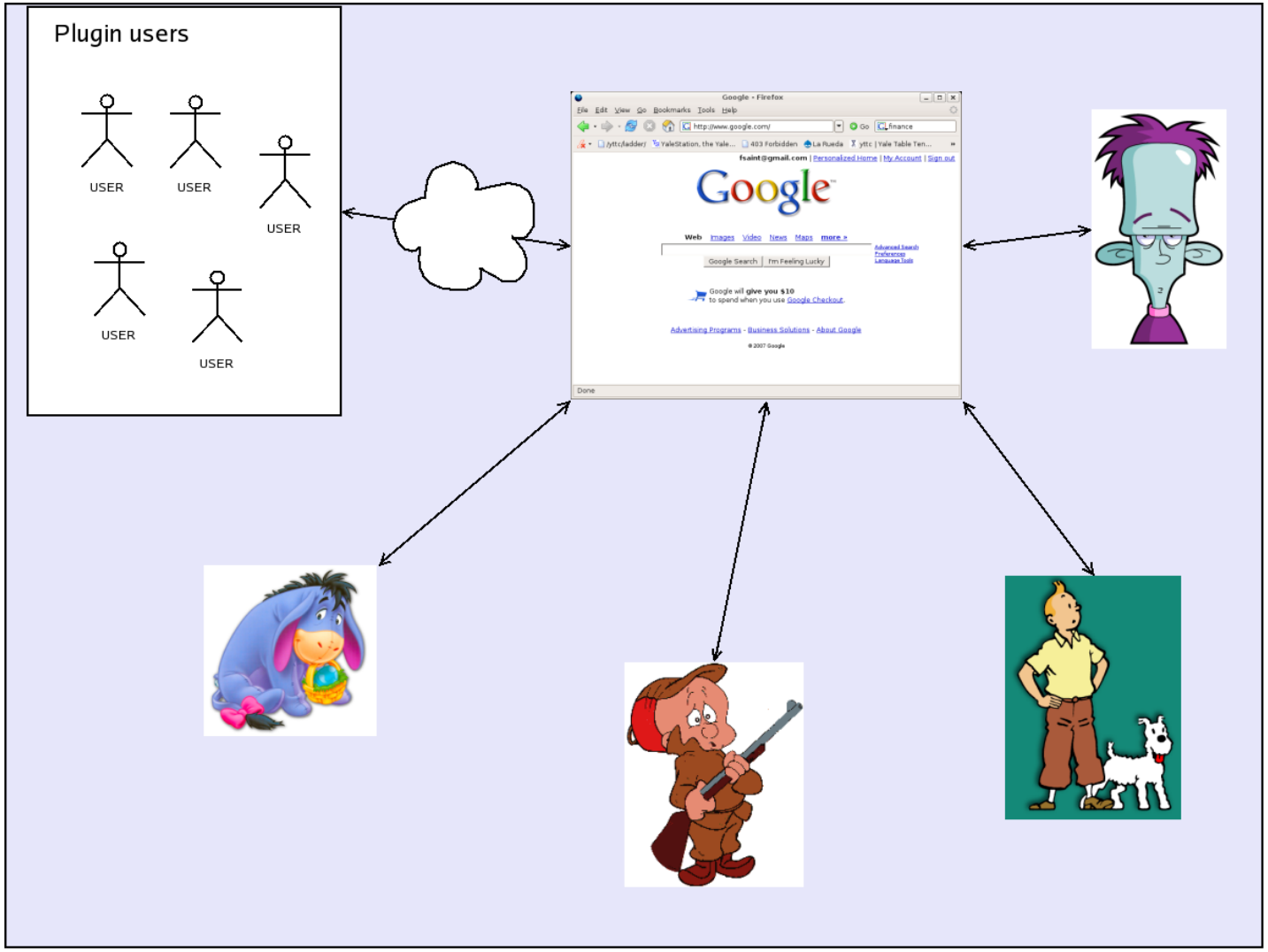
Approaches and solutions

- TrackMeNot: Firefox plugin that obfuscates real Web searches by issuing fake ones. "Cover traffic."
  - Good: Fast
  - Bad: Unclear how hard it is to distinguish real queries from fake ones. Search engine optimization is harder.
- Tor+Privoxy: General anonymous web-navigation technology.
  - Good: Tor is believed to be a good anonymity tool. Robust and stable
  - Bad: Vulnerable to active components and hard to use.
- FoxTor: Firefox plugin to manage Tor preferences. It requires Tor+Privoxy.
  - Good: More usable than Tor+Privoxy
  - Bad: Same as Tor+Privoxy

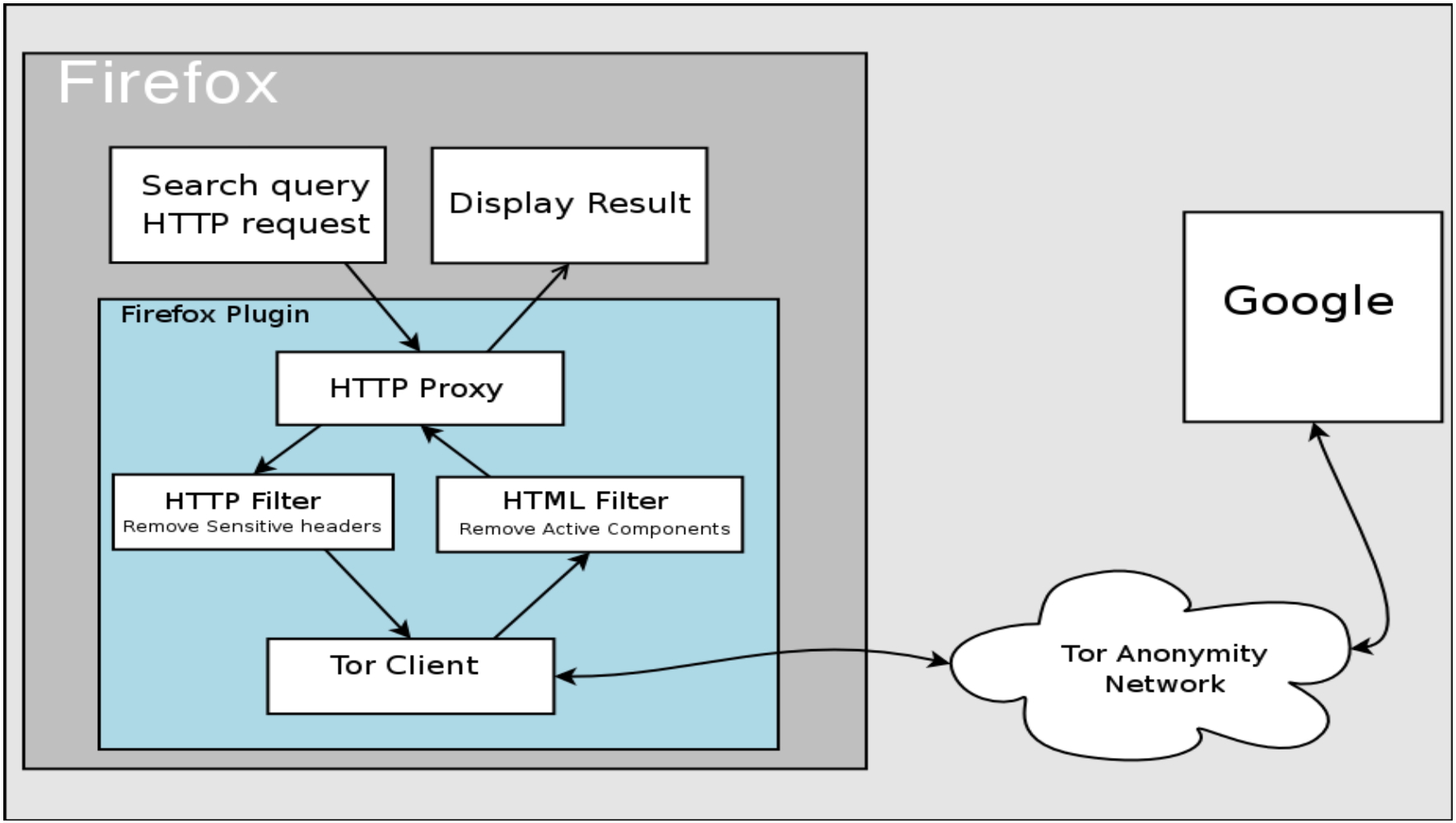
Objective: Make Users Indistinguishable (1)



Objective: Make Users Indistinguishable (2)



Design Overview



How each type of information is handled

- TCP/IP ← Tor
  - IP address
  - Institution or ISP
  - Operating System
  - uptime
- HTTP Headers ← HTTP filter
  - Cookies
  - Operating system make and version
  - Browser make and version
  - Encodings and language
- HTML ← HTML filter
  - JavaScript collected information
  - Timing information
- Query terms and time ← Can we do anything?
- Active components ← HTML filter
  - …

How each type of information is handled

- TCP/IP ← Tor
  - IP address
  - Institution or ISP
  - Operating System
  - uptime
- HTTP Headers ← HTTP filter
  - Cookies
  - Operating system make and version
  - Browser make and version
  - Encodings and language
- HTML ← HTML filter
  - JavaScript collected information
  - Timing information
- Query terms and time ← Can we do anything?
- Active components ← HTML filter
  - …

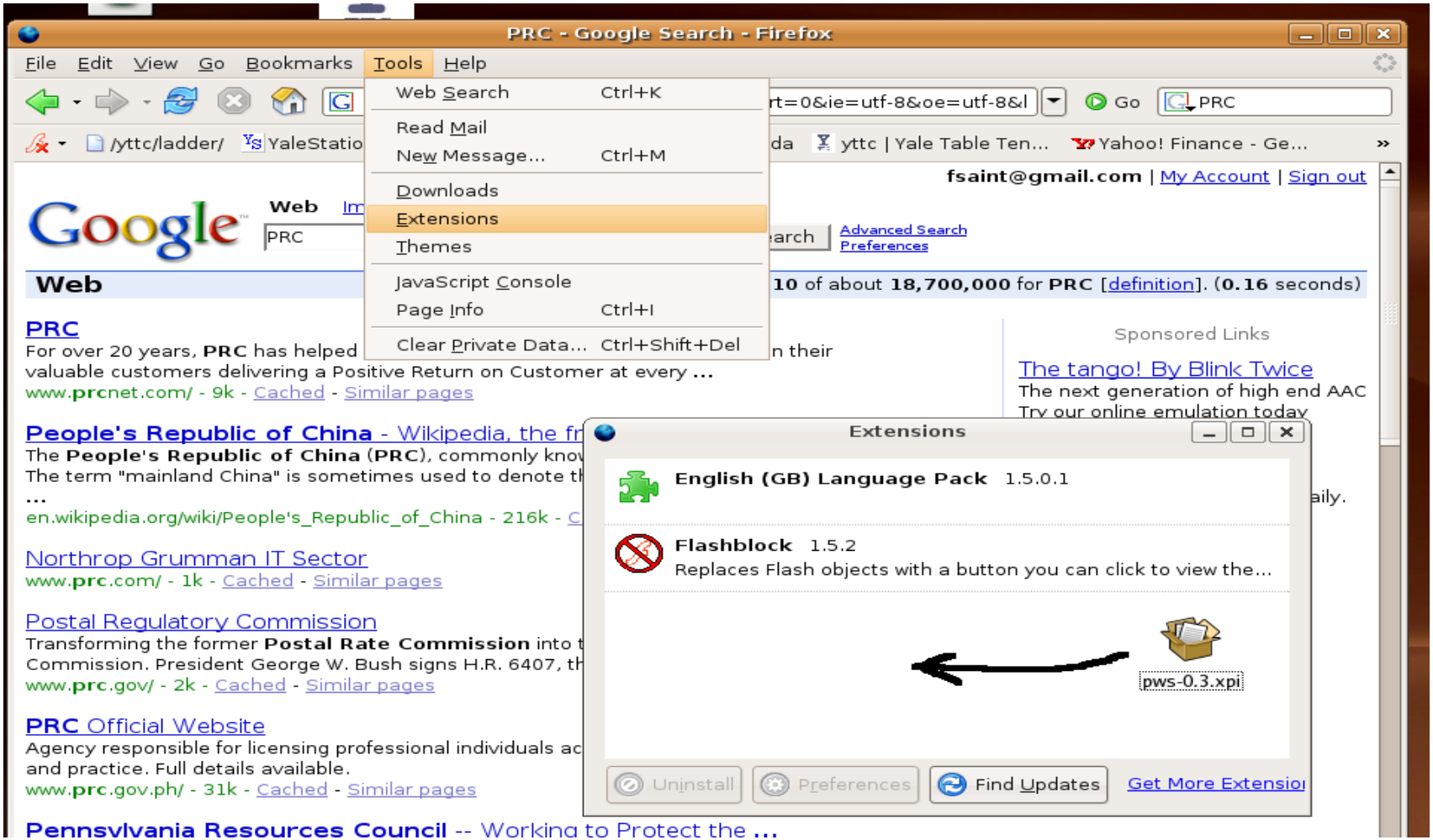
How each type of information is handled

- TCP/IP ← Tor
  - IP address
  - Institution or ISP
  - Operating System
  - uptime
- HTTP Headers ← HTTP filter
  - Cookies
  - Operating system make and version
  - Browser make and version
  - Encodings and language
- HTML ← HTML filter
  - JavaScript collected information
  - Timing information
- Query terms and time ← Can we do anything?
- Active components ← HTML filter
  - …

How each type of information is handled

- TCP/IP ← Tor
  - IP address
  - Institution or ISP
  - Operating System
  - uptime
- HTTP Headers ← HTTP filter
  - Cookies
  - Operating system make and version
  - Browser make and version
  - Encodings and language
- HTML ← HTML filter
  - JavaScript collected information
  - Timing information
- Query terms and time ← Can we do anything?
- Active components ← HTML filter
  - …

Plugin installation



Plugin use


Google - Firefox
File Edit View Go Bookmarks Tools Help
http://www.google.com/
Go
privacy concerns
Google
Yahoo
Amazon.com
Answers.com
Creative Commons
Ubuntu Package Search
eBay
Wikipedia topic
PWS Google search
Search PWS Google
Web Images Video News Maps more »
Google Search
I'm Feeling Lucky

Future Work

- Queries can still be linked at the semantic level.
- Develop a formal model to measure privacy.
- Reduce impact of Tor's path selection on performance.

- (That was six years ago. Presumably, the Tor people have made progress since then.)

The Challenge of PII in a Networked Society

JOAN FEIGENBAUM
http://www.cs.yale.edu/homes/jf/
Polytechnic Inst of NYU; Fall 2008

PORTIA's PII-related Outputs Include:

- Browser plug-ins for anonymous search
  - PWS (Private Web Search)
  - TrackMeNot
- Browser-based anti-phishing tools
  - PwdHash
  - SpoofGuard
  - SafeCache
  - SafeHistory
- Cryptographic-protocol solutions
  - MySQL PIR
  - FairPoll

JF's PORTIA Conclusions

- Less and less sensitive information is truly inaccessible. The question is the *cost* of access, and that cost is decreasing.
- Foundational legal theories to support obligations and rights in cyberspace are lacking.
- Technological progress is still going strong, almost 30 years after Diffie-Hellman, but adoption is slow.
- **Client-side defenses can only go so far.**

What's Next?

- We need a paradigm shift on PII.
- Traditional data security is based on *preventing unauthorized access* to sensitive information.
- Internet-age data security should be based on *ensuring appropriate use* of sensitive information.
- **"Hide it or lose it" won't work in a networked society. We should strive for accountability, not secrecy.**

Things Have Gotten Worse (1)

- There has been a discontinuity since this course was last taught in 2011.
- It's become less clear that there's a meaningful distinction between sensitive and non-sensitive information. Almost all information seems to be potentially useful to someone with power.
- The scale of information collection, both legal and illegal, has increased dramatically.

Things Have Gotten Worse (2)

- There is immense confusion and dread on the part of the general public about how information can be used against it – ***and those feelings are completely justified.***
- The US government has not been helpful (to put it mildly).
  - Manning // SIPRNet // Wikileaks
  - Snowden // NSA
  - CFAA // Swartz
- Probably equally true of other nat'l. govts.

This course will examine sensitive information as a nexus of technology, law, and culture.

Questions?