Is Privacy Advocacy Threatening Science? (Because of a Fear of “Big Data”)

Abe Hsuan
October 10, 2011
Explosion of data . . . but without storage capacity, data is the new “sands of time”

According to the Economist, ‘the amount of digital information increases 10X every five years outpacing available storage.’

(From “Data, data everywhere” The Economist February 27, 2010)
With Big Data, the approach to science — theorize, model, test — is changing. Whole fields are being revolutionized:

Genomics
Health Care
Social Sciences
Crime Prevention
Epidemiology
etc. etc.
Big Data poses new challenges for Privacy

Most think about “Privacy laws” as enforcement and protection against tracking

But regulators, legislators, courts are struggling with how to apply a “rule of law”
California court found plaintiffs in a class action against behavioral ad network (Specific Media) failed to articulate how use of “flash cookies” caused them economic injury.

Doubted plaintiffs could, even as a “class action,” meet the $5,000 threshold of injury required.
Not to be mixed up with security breaches . . .

<table>
<thead>
<tr>
<th>Industry</th>
<th>2009 Average Cost</th>
<th>2010 Average Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public</td>
<td>$81</td>
<td>$133</td>
</tr>
<tr>
<td>Retail</td>
<td>$133</td>
<td>$185</td>
</tr>
<tr>
<td>Media</td>
<td>$149</td>
<td>$192</td>
</tr>
<tr>
<td>Technology</td>
<td>$144</td>
<td>$236</td>
</tr>
<tr>
<td>Transportation</td>
<td>$121</td>
<td>$203</td>
</tr>
<tr>
<td>Education</td>
<td>$112</td>
<td>$237</td>
</tr>
<tr>
<td>Energy</td>
<td>$136</td>
<td>$259</td>
</tr>
<tr>
<td>Industrial</td>
<td>$159</td>
<td>$249</td>
</tr>
<tr>
<td>Consumer</td>
<td>$153</td>
<td>$192</td>
</tr>
<tr>
<td>Hospitality</td>
<td>$160</td>
<td>$256</td>
</tr>
<tr>
<td>Services</td>
<td>$209</td>
<td>$301</td>
</tr>
<tr>
<td>Financial</td>
<td>$249</td>
<td>$353</td>
</tr>
<tr>
<td>Communications</td>
<td>$294</td>
<td>$380</td>
</tr>
<tr>
<td>Healthcare</td>
<td>$301</td>
<td>$330</td>
</tr>
<tr>
<td>Pharmaceutical</td>
<td>$310</td>
<td>$345</td>
</tr>
</tbody>
</table>

Ponema Institute & Symantec 2010 Study on U.S. Cost of Data Breach
Privacy is amorphous, multifaceted

Common framework looks at privacy as protection against:

1. Intrusion upon private affairs
2. Public disclosure of embarrassing facts
3. Publicity in a false light
4. Appropriation of name or likeness for someone else’s advantage
Harm has to be “legally cognizable,” but these days what is being protected against is often intangible

California Supreme Court stops retailer from asking for “zip code” when customers use a credit card. According to the lawyers explaining their case:

“A lot of the people we talked to felt very uncomfortable giving the ZIP Code but felt they had to.”
Prof. Latanya Sweeney’s study of 1990 U.S. Census Data famously raised the bar for privacy to cover re-identification.

Analysis of the 1990 U.S. Census data showed that 87% of the U.S. Population could be identified uniquely by three non-PII data points:

- Gender
- Birth Date
- 5-digit Zip Code (U.S. Provincial code; e.g., “Beverly Hills 90210”)

Replicated in 2005 study (61% based on 1990 Census and 63% based on 2000 Census)
Sweeney picked an example to demonstrate

- Massachusetts Group Insurance Commission (GIC) released state employees’ records *for research purposes.*
- GIC removed name, address, social security number, and other “explicit identifiers” but included zip code, gender and birth date.
- Massachusetts Governor William Weld assured the public that GIC had protected patient privacy by deleting all PII.
She purchases voter rolls for Cambridge, Mass for US$20, a public database of name, address, ZIP code, birth date, and gender of every voter.

6 people in Cambridge shared Weld’s birth date, only 3 were men, and only one lived in his ZIP code.

Thus, Sweeney discovered Governor Weld’s health records (diagnoses and prescriptions).
HIPAA De-Identification Requirements Revised

- names
- **geo subdivisions** (smaller than a state)
- **all elements of dates** (except year)
  - including dates of admission, discharge, birth, death and, for individuals (over 89 years old, the year of birth must not be used)
- telephone and fax
- Email addresses
- Government-issued ID (SSN, driver license, etc.)
- medical record numbers
- health plan numbers
- account numbers
- certificate/license numbers
- vehicle identifiers
- device identifiers (serial #)
- web URLs & IP Addresses
- biometric identifiers (e.g., finger, voice prints)
- full face photos
- Other unique ID number, etc.
Privacy Advocates arguing all “Big Data” (not just health data) should live up to this level:

- No longer sufficient that data sets be “anonymized”

- They must be “de-identified” sufficiently to prevent re-identification
  - HIPAA requires an opinion from a qualified statistical expert that the risk of identifying an individual is very small under the circumstances; the methods and justification for the opinion should be documented.
Challenge is that world is increasingly connected, making it easier to re-identify:
In October 2006, Netflix released de-identified user records for 100 million records of over 500,000 users spanning 1999 to 2005.

- Each record has movie title, 1-5 star movie rating, and date of rating.
- Records represent less than $1/10^{th}$ of Netflix data in 2005.
Turns out the dataset is very sparse . . .

• 99% of subscribers can be uniquely identified by knowing 8 movie ratings (with 2 wrong) and dates of ratings (with a 14-day error).

• 68% based on two ratings and dates (with a 3-day error)

Figure 4. Adversary knows exact ratings and approximate dates.

From: Arvind Narayanan and Vitaly Shmatikov, How to Break the Anonymity of the Netflix Prize Dataset (October 16, 2006)
Even knowing only the movies rated, without the date of ratings . . .

84% of subscribers can be uniquely identified if 6 out of 8 “long tail” movies (i.e., outside the Top 100/500) “blockbusters”

Figure 8. Adversary knows exact ratings but does not know dates at all.

<table>
<thead>
<tr>
<th>Not in X most rated</th>
<th>% of subscribers who rated ...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≥ 1 movie</td>
</tr>
<tr>
<td>$X = 100$</td>
<td>100%</td>
</tr>
<tr>
<td>$X = 500$</td>
<td>99%</td>
</tr>
<tr>
<td>$X = 1000$</td>
<td>97%</td>
</tr>
</tbody>
</table>
Netflix – IMDb mashup

It took Univ of Texas professors Arvind Narayanan and Vitaly Shmatikov just 2 weeks to match Netflix records with public information scraped from IMDb.com to identify 2 subscribers:
Netflix “blunder” is one of the most famously cited examples of a “data valdez” … but is it?

Netflix is about to commit a privacy Valdez with its customers' viewing data

By Cory Doctorow at 12:05 AM Tuesday, Sep 22

Princeton’s CU Boulder’s Paul Ohm writes about Netflix's huge privacy blunder. A million customers have had their most sensitive viewing information -- ZIP code, age, gender, birth and death, gender and ZIP -- paraded across the internet. Ohm says Netflix is about to purchase Americans into the privacy Valdez.

Freedom to Tinker

... is your freedom to understand, discuss, repair, and modify the technology

Netflix has announced plans to commit a privacy blunder that could cost it millions of dollars in fines and civil damages.

In my last post, I had promised to say more about my article on the limits of anonymization and the power of reidentification. Although I haven’t said anything for a few weeks, others have, and I especially appreciate posts by Susannah Fox, Seth Schoen, and Nate Anderson. Not only have these people summarized my article well, they have also added a lot of insightful commentary, and I commend these three posts to you.

Today brings news relating to one of the central examples in my paper: Netflix has announced plans to commit a privacy blunder that could cost it millions of dollars in fines and civil damages.

In my article, I focus on Netflix’s 2006 decision to release millions of records containing the movie rating preferences of “anonymized” users to the public, in order to fuel a crowd-sourcing competition called the Netflix Prize. The Netflix Prize has been a huge win
Typical “data valdez” is a lost laptop or security hole but this Aol mishap expanded the definition:

In August 2006, America Online publicly posted:

-- 20 million search queries
-- sample of 650,000 users
-- covering three months of activity

AOL had deleted member data, IP addresses, etc.

AOL released the search histories to “embrace the vision of open research”
BUT Aol failed to “clean” the search queries of PII:


Arnold acknowledged her searches which included “mildly embarrassing” queries such as “numb fingers”, “60 single men”, and “dog that urinates on everything.”
Netflix is not a “data valdez” in the normally understood sense … still this is the result …

**NetFlix Cancels Recommendation Contest After Privacy Lawsuit**

By Ryan Singel  March 12, 2010 | 2:48 pm | Categories: privacy

Netflix is canceling its second $1 million Netflix Prize to settle a legal challenge that it breached customer privacy as part of the first contest’s race for a better movie-recommendation engine.

Friday’s announcement came five months after Netflix had announced a successor to its algorithm-improvement contest. The company at the time said it intended to expand the amount of information it gave to researchers in hopes that its recommendation system — a key part of Netflix’s customer retention strategy — would get even better. That was then followed with a warning by prominent data privacy lawyers that the new dataset was easily de-anonymized.

Those fears were highlighted in December, when an in-the-closet lesbian mother sued Netflix for privacy invasion, alleging the movie-rental company made it possible for her to be outed when it disclosed insufficiently anonymous information about nearly half-a-million customers as part of its $1 million contest.
“Data Valdez threat real?”: Cheap access, storage and processing costs mean anyone can play...
And, compared to the past, data can be processed much (much!) faster … which adds to the anxiety:

Took Donnelly, Polk, etc. **5 years** to mash 1970s U.S. Census data (aggregates of 1,500 households) into datasets that were demographically accurate for over 50% of all households.
With networked PCs and mobile devices, publicly available “personal data” is growing exponentially
Where all the data lives … many operating many times removed from user’s actual point of contact
Implicit message is that anyone and everyone exposed by simply how they use the Net.

How much responsibility should people take for re-identification is based on their own “self” publication or disclosures?
Professor Paul Ohm speaks of a “Database of Ruin”

In “Broken Promise of Privacy,”* an influential paper by leading privacy advocate and law professor Paul Ohm:

“For almost every person on earth, there is at least one fact about them stored in a computer database that an adversary could use to blackmail, discriminate against, harass, or steal the identity of him or her.”

For almost every one of us, then, we can assume a hypothetical “database of ruin”

• Prosecute inadequate anonymization efforts
• Advocate total ‘opt-in’ consent model*
• Suspicious (paternalistic) view of social media (protect users against themselves)

Result: Chilling effect on all kinds of uses of data, including R&D

Harvard Researchers Accused of Breaching Students' Privacy

Social-network project shows promise and peril of doing social science online

By Marc Parry

In 2006, Harvard sociologists struck a mother lode of social-science data, offering a new way to answer big questions about how race and cultural tastes affect relationships.

The source: some 1,700 Facebook profiles, downloaded from an entire class of students at an "anonymous" university, that could reveal how friendships and interests evolve over time.

It was the kind of collection that hundreds of scholars would find interesting. And in 2008, the Harvard team began to realize that...
Fine line: Zealous protection of privacy can lead to censorship or arbitrary prosecution
Library of Congress, Facing Privacy Concerns, Clarifies Twitter Archive Plan

May 7, 2010, 12:00 pm
By Marc Parry

If you don’t want that tipsy 3 a.m. Twitter post preserved for posterity, then start deleting. Now.

Faced with privacy concerns, the Library of Congress is clarifying its plans to archive all public tweets posted since Twitter went live in March 2006. The database won’t contain deleted tweets or private account information, according to a list of frequently asked questions recently posted on the library’s blog.

And the Twitter database will only be made available to “qualified researchers,” Martha Anderson, director of the library’s National Digital Information Infrastructure and Preservation Program, tells The New York Times. The plan is to embargo messages for six months before making them available, but that period could be extended, she says.

“There’s concern about privacy issues in the near term, and we’re sensitive to these concerns,” she says. “We may have to filter certain things or wait longer to make them
IMS Health – Provides data-mined prescription drug history data to pharmaceutical companies

MEDICARE PART D AT AGE FIVE

Senior Rx Drug Prices: Steadily Downward

Cost declines observed in the first five years of Medicare Part D, fueled in part by cheaper generic drugs, will likely continue through 2015, says an IMS Institute for Healthcare Informatics study.

Download the Report
June 2011 U.S. Supreme Court ruling which affirmed, on free speech grounds, the right of pharmaceutical marketing companies to data-mine personally-identifiable prescription information and sell the resulting marketing information to pharmaceutical companies

On first blush . . . .
Many boundaries are dissolving and “old rules” no longer apply
“No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks.”

Adopted by the United Nations General Assembly on December 10, 1948
“No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against arbitrary interference or attacks.”

Adopted by the United Nations General Assembly on December 10, 1948
Legal definition of “arbitrary” ... despite facial objectivity, still defined by societal views & norms

“not based on reason or judgment but on personal will or discretion without regard to rules or standards.”*

“founded on prejudice or preference rather than on reason or fact.”**

“implies an element of bad faith, and it may be used synonymously with tyrannical or despotic.”*

In America, the “right of privacy” born when this was invented ... snapshot photography ... in 1888

Snapshot cameras altered our sense of when we should have a reasonable expectation of privacy . . . a “right to be left alone”

First “privacy” law enacted by NY in 1903 mostly to address uses of photographic likenesses
Frequency of the words "secrecy" versus "privacy" in books from 1890-2000
“Privacy” tends to come to fore when there is a discomfort with what technology makes possible.
More than 30 years later, many people, even in Europe, still believed the camera could steal one’s soul.

This from James Frazer’s *The Golden Bough* (published in 1922):

“There are persons in the West of Scotland who refuse to have their likenesses taken lest it prove unlucky; and give as instances the cases of several of their friends who never had a day’s health after being photographed.”
Are we merely talking about another kind of “photograph” – that of our “digital shadow”? 

http://digitalslander.files.wordpress.com/2008/08/digital-shadow.jpg
In the meantime, science is the biggest victim of a potential “Tragedy of Data Commons”

With privacy advocates scaring consumer and legislators, Brooklyn Law Professor Jane Yakowitz says we are heading towards a “tragedy of the data commons” effect?

Users opt out of adding data to the commons despite benefiting, until the collective benefit degenerates.

“Technology will move faster than governments, so don't legislate before you understand the consequences.”

Eric Schmidt at G8
When (Leg)islators run too far ahead…
Thank You

Abe Hsuan
AbeHsuanEsq@gmail.com