ifis

Institut für Informationssysteme
Technische Universität Braunschweig

Frisst die Informationstechnik ihre Kinder?

May 9, 2016

Wolf-Tilo Balke

Roadmap

Some thoughts about information technology…

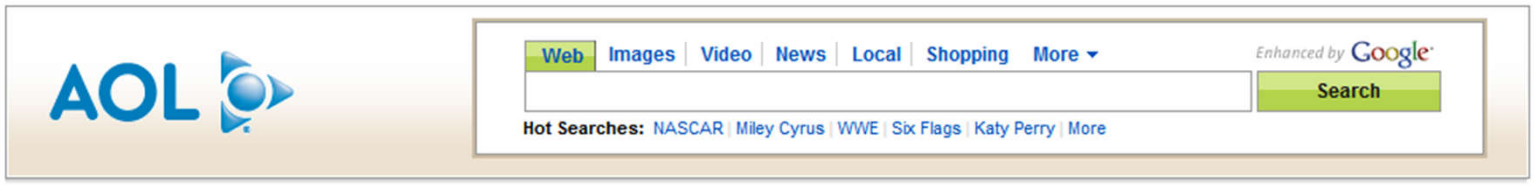
- Web Search Engines
  - Dangerous gateways to the world
- Classic statistical data security
  - Dual use and limits of anonymization
- Connecting the dots in information networks
  - Information fusion
- What do we know outside of the network?
  - Indirect information fusion

Web Search Engines

- Today, Web Search Engines are our de-facto gateways to the world
  - The Web of Information: Wikipedia, YouTube, Google Maps…
  - Google's co-founder Larry Page: "The perfect search engine would understand exactly what you mean and give back exactly what you want."
  - But this is also shaping our view of the world: filter bubbles & echo chambers
  - And… who is in control?
    - N. Kulathuramaiyer and W. - T. Balke, "Restricting the View and Connecting the Dots - Dangers of a Web Search Engine Monopoly", Journal of Universal Computer Science (J.UCS), vol. 12, no. 12, 2006.

Your data is safe with us…

- The tale of the anonymous dataset: **AOL Search**
  - One of the major web search and content portals
  - AOL serves millions of searches per day
- Of course AOL has a **privacy policy…**
  - However, internally **records** are kept of all user searches
  - Search records are **very valuable** for improving algorithms
  - In 2006, **search data** of 650,000 users over a 3-month period was published for free use by the IR research community



Your data is safe with us…

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

The New York Times

PRINT
SINGLE PAGE
REPRINTS

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

Erik S. Lesser for The New York Times
Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

What Does Search Data Tell?

- Most prominent example: User #**4417749**
  - Thelma Arnold, 62-year-old, widowed, lives in Lilburn, Georgia
  - Is looking for a new partner in his 60s
  - Has at least one dog randomly pissing on furniture
  - Has problem with trembling fingers and aches in her back
  - Is worried about the safety of her neighborhood
  - Wonders about problems of the world, like hunger in Africa or children in war-torn Iraq

What Does Search Data Tell?

- User 311045:
  - how to change **brake pads** on **scion xb**
  - 2005 us open cup **florida** state champions
  - how to get revenge on a ex
  - how to get revenge on a ex girlfriend
  - how to get revenge on a friend who f---ed you over
  - **replacement bumper** for scion xb
  - **florida** department of law enforcement
  - crime stoppers **florida**

What Does Search Data Tell?

- User 11574916:
  - **cocaine** in urine
  - asian mail order brides
  - states reciprocity with **florida**
  - florida dui laws
  - **extradtion from new york to florida**
  - mail order brides from largos
  - will one be extradited for a **dui**
  - **cooking jobs** in french quarter new orleans
  - will i be extradited from ny to fl on a dui charge

Limits of Anonymization

Commonwealth of Massachusetts
Group Insurance Commission

- The **Massachusetts Group Insurance Commission** had a bright idea back in the mid-1990s - it decided to release "anonymized" data on state employees
  - The data showed all hospital visits after removing all obvious identifiers such as name, address, and social security number
  - At the time of release, **William Weld**, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting all identifiers

Limits of Anonymization

- Latanya Sweeney (Harvard U) started hunting for the Governor's hospital records in the GIC data
  - Governor Weld resided in Cambridge, MA, a city of 54,000 residents and seven ZIP codes
    - For twenty dollars, she purchased the complete voter rolls from the city, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter
  - By combining this data with the GIC records, Sweeney found Governor Weld with ease
    - Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code
  - Finally she sent the Governor's health records (which included diagnoses and prescriptions) to his office.

Limits of Anonymization

- De-Anonymization usually works amazingly well not only in medical contexts…
  - 87% of all Americans could be uniquely identified using only three pieces of information: ZIP code, birthdate, and sex
    - Sweeney, L. A. *Simple Demographics Often Identify People Uniquely*. Carnegie Mellon Univ. Data Privacy Working Paper 3 (2000).
  - Failures of anonymization in system interactions are not really understood by users
    - Ohm, P. Broken promises of privacy: responding to the surprising failure of anonymization. UCLA Law Rev. 57, 1701 (2010).

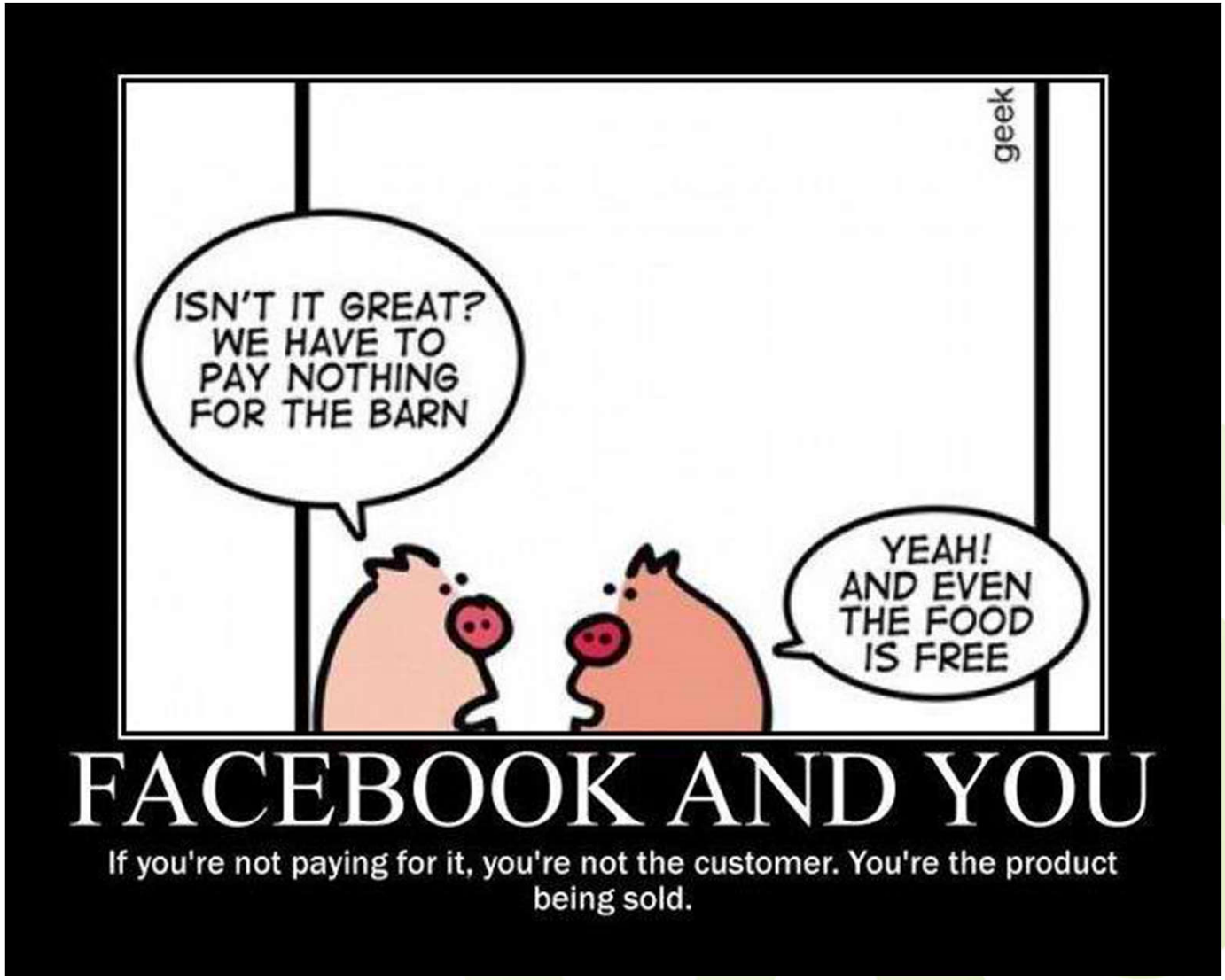
Information Fusion

- Even worse… Social Networks
  - Every **interaction** with social networks, every **piece** of information, every **decision** for accepting/rejecting a friendship request tells something about the user and can be used as an attribute for later **classification**
    - What Facebook „likes“ tell about members
      Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. PNAS 110 (15).
    - What Facebook friendships tell about sexual orientation
      Jernigan C, Mistree BFT (2009) Gaydar: Facebook Friendships Expose Sexual Orientation. First Monday14(10).

Indirect Information Fusion

- Indirectly gathered information is the most unintuitive thing in information fusion
  - When a user joins a social network, he/she will explore the environments
    - "Who of my real world friends is on this network?"
    - The social network provider will not forget these names…
  - What Facebook knows about non-member friends of members
    - Horvát E-Á, Hanselmann M, Hamprecht FA, Zweig KA (2012) One Plus One Makes Three (for Social Networks). PLoS ONE 7(4): e34740.

To Get The Discussion Started….



Thanks for the Attention!

Technische Universität Braunschweig