Gregory Conti
United States Military Academy
West Point, New York
gregory-conti@usma.edu
The views expressed in this presentation are those of the author and do not reflect the official policy or position of the United States Military Academy, the Department of the Army, the Department of Defense or the U.S. Government.
Scenarios

• Patent search
• Serious medical condition
• Failing company
• Anonymous blogger
• Death in the family
Usama Fayyad
Chief Data Officer
2005 Interview with ACM

http://www.acm.org/sigs/sigkdd/explorations/issues/7-2-2005-12/fayyad.html
What are some of the biggest data mining challenges you face now at Yahoo?

“Yahoo!'s users, through their use of our network of products, generate over 10 terabytes of data per day. This is the equivalent of the entire text contents of the library of Congress. This is data that describes product usage, and does not include content, email, or images, etc.”

“The first and largest challenge is the ability to capture all of this data reliably, process it, reduce it, and use it to feed the many, many [reports and applications]”

http://www.acm.org/sigs/sigkdd/explorations/issues/7-2-2005-12/fayyad.html
What about Privacy and Data Mining at Yahoo?

“Yahoo! was built around consumer trust and it has been our number one priority since day one. We would never do anything to compromise our users privacy.”

“Yahoo!, perhaps more than any other online consumer company, understands that the long-term survival and health of the business is built on consumer trust.”

http://www.acm.org/sigs/sigkdd/explorations/issues/7-2-2005-12/fayyad.html
Organize the world's information and make it universally accessible and useful.

Evil
Generate revenue by providing advertisers with the opportunity to deliver measurable, cost-effective online advertising

Good

http://www.nytimes.com/2005/10/30/business/yourmoney/30google.html?ex=1288324800&en=b0684c6ec54b2467&ei=5090&partner=rssuserland&emc=rss&pagewanted=all
"Google.cn granted license for operation in China"
censorship required to do "legal business" in China

"American Airlines subpoenas Google, YouTube"
AA demands Google reveal name of person who posted airline's training videos.

"Google Subpoena Woes Double"
DOJ, Child Pornography

"Yahoo helped Jail China writer"

"Phishing with Google Desktop"
(remote code execution via malicious website)

Sony Rootkit

Google balances privacy, reach
Elinor Mills (CNET)
What if an entirely altruistic provider

- Insider threat
- Accidental information disclosure
- Merger / Sale of Company
- Change in leadership / corporate philosophy
- Impact of technical advances
- Legal compulsion

What if not?
Information Disclosed

URL Shortening Service
Online Auction
Instant Messaging Provider
Web Search
Free Email Provider
File Transfer Service
Consolidated Information Service Provider

Tiny URL
16 Million URLs
325M hits/month
I have searched for things I wouldn’t want my grandmother* to know about.
Case Study: Google
Google Zeitgeist 2005

Google.com
Top Gainers of 2005
1. Myspace
2. Ares
3. Baidu
4. wikipedia
5. orkut
6. iTunes
7. Sky News
8. World of Warcraft
9. Green Day
10. Leonardo da Vinci

Google News
Top Searches in 2005
1. Janet Jackson
2. Hurricane Katrina
3. tsunami
4. xbox 360
5. Brad Pitt
6. Michael Jackson
7. American Idol
8. Britney Spears
9. Angelina Jolie
10. Harry Potter

Froogle
Top Searches in 2005
1. ipod
2. digital camera
3. mp3 player
4. ipod mini
5. psp
6. laptop
7. xbox
8. ipod shuffle
9. computer desk
10. ipod nano
Google Talent

“Passionate about these topics? You should work at Google.”

- algorithms
- artificial intelligence
- compiler optimization
- computer architecture
- computer graphics
- data compression
- data mining
- file system design
- genetic algorithms
- information retrieval
- machine learning
- natural language processing
- operating systems
- profiling
- robotics
- text processing
- user interface design
- web information retrieval
- and more!

http://labs.google.com/
Information you provide or retrieve

↑ ↑ ↑

Information others provide or retrieve

Google

Information Google retrieves (think Googlebot)
Disclosure Vectors
“When Did You Lose Your Google Virginity?”

Quote by Scott Lemon, http://the.inevitable.org/anism/
Information you provide or retrieve

Information others provide or retrieve

Information Google retrieves (think Googlebot)
Google Research...

google
google api
google autolink
google base
google blacklist
google chat
google contest
google hacked
google intervened
google isp
google keyhole
google labs
google maps
google scholar
google screen
google sets
google toolbar
google wallet
google watch
google wireless
googlebot
I use Google as an Address Book.
Google Alerts

- search terms
- email address
- frequency
- category (news, web, news and web, groups)

http://www.google.com/alerts
Google Maps / Satellite Imagery

Locations of interest to you (down to street level)

What have you looked at?
Gmail

• email recipients
• email sources
• content
• N-order contact
Gmail

- email recipients
- email sources
- content
- N-order contact
Gmail

- email recipients
- email sources
- content
- N-order contact
I use a Gmail account.
I have sent an email to a Gmail account.
My Employees are registered Gmail users.
Residential/Commercial Phone Number Lookup

a search returns...

- name
- phone number
- address
- links to 3 mapping services
Travel Support

Google Search - Mozilla Firefox

Web Images Groups News Froogle Local

Search for lax nyc

Results 1 - 10 of about 4,160,000 for lax nyc. (0.12 seconds)

Flights from Los Angeles, CA to New York, NY

Departing: 11/24 Returning: 12/01

Search: Expedia Hotwire Orbitz Priceline

Local results for lax near New York, NY

Robert I Lax & Assoc - 3.1 miles NE - 535 5th Ave # 21, New York, 10017 - (212) 818-9150
Lax Martin - 3.2 miles N - 1185 Avenue of the Americ, New York, 10036 - (212) 372-1208
Lax Moshe - 3.4 miles NE - 250 E 49th St, New York, 10017 - (212) 758-9840

LAX NYC Cheap Airfare

cheap airfare to NYC from LAX to New York City from Los Angeles.

WDPCC | Teams Contacts

Orca Lacrosse NYC New York, NY Philadelphia Premier Philadelphia, PA Philly Women's
Lax Philadelphia PA Portland Women Portland, OR
Collaborative Word Processing

The Web Word Processor
...that's now part of Google.

- Share documents instantly & collaborate real-time.
  Pick exactly who can access your documents.
- Edit your documents from anywhere.
  Nothing to download -- your browser is all you need.
- Store your documents securely online.
  Offline storage plus data backup every 10 seconds.
- Easy to use.
  Clean, uncluttered screens with a familiar, desktop feel.

Let me know when I can sign up!

E-mail: [ ]
Submit

We have closed off new registrations until we move Writely to Google's systems.

Please submit your e-mail address and we'll send you an email when registration opens back up.

→ Take a tour → Find out more
Can I ... upload from Word? YES!
Can I ... save to my desktop? YES!
Can I ... publish on the web? YES!
Can I ... post on my blog? YES!
Google Calendar
Local Information for Local Devices

- Local for Mobile
  - detailed directions
  - search results integrated with map
  - zoom in/out, drag maps
  - satellite imagery

- free download

http://www.google.com/glm/index.html
Google Desktop

“Info when you want it, right on your desktop”

http://desktop.google.com/
Search Appliances

The Google Search Appliance makes the sea of lost data on your web servers, file systems and relational databases instantly available with one mouse click.

http://www.google.com/enterprise/gsa/index.html
Information you provide or retrieve

Information others provide or retrieve

Information Google retrieves (think Googlebot)
Keyphrases
(from Google.com)

• greg conti
• gregory conti
• a framework for countering denial-of-information attacks
• ids rainstorm
• network security visualization
• 124th mi bn
• awareness teach information security filetype:pdf
• passive visual fingerprinting of network attack tools
• usma ia
• 92d information warfare aggressor squadron
• conti gatech
• greg conti rumint
• hacking and innovation cacm
Information others provide or retrieve

Information you provide or retrieve

Information Google retrieves
(think Googlebot)
Fingerprinting
Service Provider A

Information Disclosed

Time

Fingerprinting Threshold A

Service Provider A

Time

Information Disclosed

Fingerprinting Threshold A

$F_A$
Groups
Instant Messaging
Mapping
Web Searches
Translation
Web Searches
Email
Finance
Text Search
Scholar
Local
Book Search
Calendar
Shopping
News
Groups
Blog
Mapping
Alerts
Language Translation
Word Processing
Network Layer
Cookies
Browser Environment Variables
Registration
Semantics
Work PC

Network Layer
Cookies
Browser Environment Variables
Registration
Semantics
Network Layer
Cookies
Browser Environment Variables
Registration
Semantics
Network Layer
Cookies
Browser Environment Variables
Registration
Semantics
I have a Google cookie on my computer.
My Employees have Google cookies on their computers.
I have registered for a Google account.
Countermeasures
Information Disclosed A

Fingerprinting Threshold A

Time

Information Disclosed A

Information Disclosed’ A
Countermeasures

- Anonymous browsing (Tor, anonymizer...)
- Diversity
- Policy/Law
- Go directly to the website you want
- Anonymizer-ish application for individuals and organizations but operates higher up the semantic scale
- Aggregators
- Encrypt content
- Switching Proxy Plug-ins
- Cookie rewriters
Countermeasures

- Multiple email accounts, host own email
- Chaffing
- DHCP/NAT
- Retain only the "necessary" amount of information
- Time limits
- Transparency
- Opt-in / out
I actively manage/remove cookies.
I regularly use web anonymization tools.
Examples
Co-worker example #1

- 696 entries

- 18 Feb 05 - 26 Jul 06
  - 1.33 unique searches / day

- Sanitization criteria
  - names of family, close friends
  - addresses
  - credit card numbers
  - SSN’s
Co-worker example #2

- 1264 entries,
- 9 Nov 05 - 31 Jul 06
  - 4.8 unique searches / day
- criteria
  - names
  - locations close to home
  - phone numbers
  - friends blog

24%
sanitization
Co-worker example #3

• 3974 entries,
• 8 Dec 04 - 31 Jul 06
  – 6.6 unique searches / day
• criteria
  • names
  • friends
  • phone numbers
  • co-workers

31% sanitization
Demo
Discussion