# Network technology: privacy implications

Keunwoo Lee
590T (Society and Technology seminar)
8 May 2006

## Plan

- What do
  - my computer & local network
  - my Internet connection
  - websites I visit

  know about me?
- How can this data be aggregated?
- How can we mitigate the risk of exposing "too much information"?

## Your computer

"Personal information":

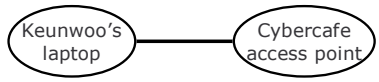
- Your files, keystrokes, etc…
- Software IDs from online registration, etc.
- Assorted hardware IDs
  - Intel unique CPU ID, hard drive serial number, etc.

No "good reason" to transmit the above; hence, you "trust" your software not to send over the network

Then there's your network card…

## Your network card

- Most *local* network hardware is Ethernet
- MAC address:
  - Every Ethernet card in the world has unique ID number called a MAC address
  - **Implicitly & necessarily broadcast to peers** whenever connecting to network (and sometimes when not)

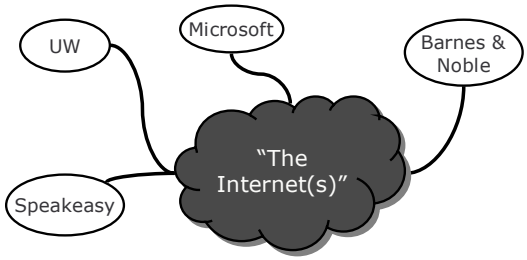



## Plan

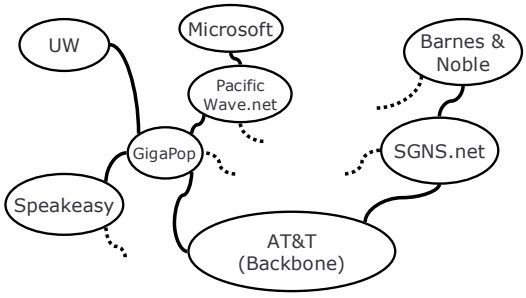
- What do
  - my computer & local network
  - **my Internet connection**
  - websites I visit

  know about me?
- How can this data be aggregated?
- How can we mitigate the risk of exposing "too much information"?

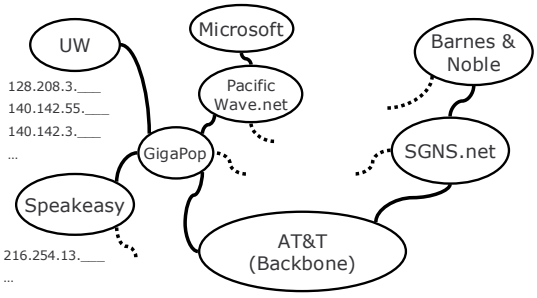
## The Internet



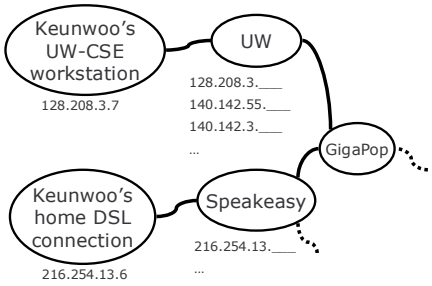
## The Internet




## Internet Protocol (IP) subnets



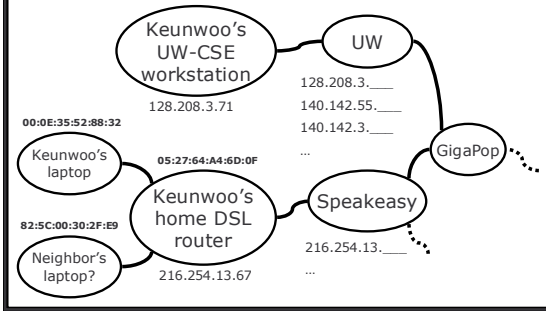

## IP address assignment



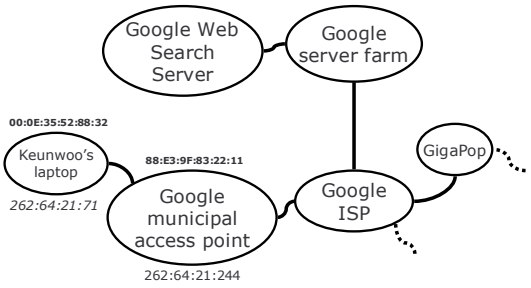

## IP address assignment

- Broadband: typically **static**, i.e. assigned for months or years on end
- Dial-up, cybercafes, other transient connections: **dynamic**, i.e. assigned for minutes/hours from pool
  - Provider can still keep logs of which customer had which IP at any given time; will produce if subpoenaed etc.
- Network address translation (NAT)
  - multiple machines to share one IP address
  - common within homes
  - plausible deniability?

## IP sharing via NAT




## When Google's your ISP…



## So what?

- IP address is *necessarily communicated* to any machine that you talk to *directly*
  - Visit website, incl. search engine
  - Send instant message
  - Share file via peer-to-peer
  - Play online game
  - …
- Many applications *log* these addresses…

## Plan

- What do
  - my computer & local network
  - my Internet connection
  - **websites I visit**

  know about me?
- How can this data be aggregated?
- How can we mitigate the risk of exposing "too much information"?

## Application protocols

- TCP/IP provides "pipes"; application protocols determine what goes through the pipes
- Email
- Web (HTTP)
- …

## HTTP

- How a browser asks a server for information
- Like all other direct Internet connections, communicates your IP address
- HTTP referer: When you click a hyperlink, your browser tells the *target* web server what page you're coming from
  - Not required, but all browsers do this by default
- Cookies: a way for web servers to *ask* your browser to store a small amount of information on their behalf
  - Browser may reject cookies

## HTTP server logs

Whenever you visit a web server (e.g., www.washington.edu), that server probably records *at least* the following:

- Your IP address
- What web browser you're using
- What language your web browser's configured to use
- The time
- The name of the page you requested

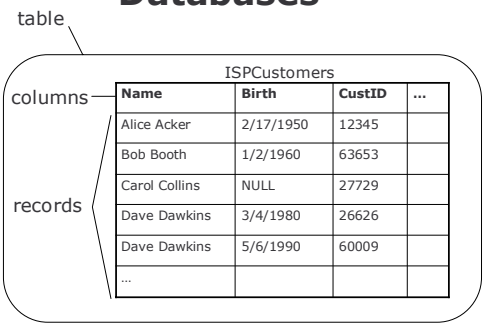
## Modern web pages: many pieces

# Plan

- What do
  - my computer & local network
  - my Internet connection
  - websites I visit

  know about me?
- **How can this data be aggregated?**
- How can we mitigate the risk of exposing "too much information"?

# Databases




| Name | Birth | CustID | ... |
|---|---|---|---|
| Alice Acker | 2/17/1950 | 12345 | |
| Bob Booth | 1/2/1960 | 63653 | |
| Carol Collins | NULL | 27729 | |
| Dave Dawkins | 3/4/1980 | 26626 | |
| Dave Dawkins | 5/6/1990 | 60009 | |
| ... | | | |

# Database fusion

- Database **join**: fundamental operation, as old as databases
- Combines records from 2 or more tables that "match" on some column value

# Join example

ISPCustomers

| Name | Birth | CustID | IPAddr |
|---|---|---|---|
| Alice Acker | 2/17/1950 | 12345 | 1.2.3.4 |
| Bob Booth | 1/2/1960 | 63653 | 1.2.3.5 |
| Carol Collins | NULL | 27729 | 1.2.3.6 |
| Dave Dawkins | 3/4/1980 | 26626 | 1.2.3.7 |
| Dave Dawkins | 5/6/1990 | 60009 | 1.2.3.8 |
| ... | | | |

TelemarketingList

| Name | Address | Merchants |
|---|---|---|
| Alice Acker | 1 First St.,... | QFC,GAP,... |
| Bob Booth | NULL | STDClinic,... |
| Dave Dawkins | 3 Third St.,... | PornGalore,... |
| ... | | |

join on **Name**, selecting Name, IPAddr, Address, Merchants

| Name | IPAddr | Address | Merchants |
|---|---|---|---|
| Alice Acker | 1.2.3.4 | 1 First St.,... | QFC,GAP,... |
| Bob Booth | 1.2.3.5 | NULL | STDClinic,... |
| Dave Dawkins | 1.2.3.7 | 3 Third St.,... | PornGalore,... |
| Dave Dawkins | 1.2.3.8 | 3 Third St.,... | PornGalore,... |
| ... | | | |

# Complications in practice

- How did one entity get both of these databases?
  - Incentives to share data?
- Hard if data doesn't match perfectly
  - What if one database used the name "David F. Dawkins"?
- Ongoing CS research problem: database fusion with imperfectly matching data
  - State of the art: can get statistically good matches, but not absolute confidence
  - What you can expect in the future (Keunwoo's non-specialist opinion): automated statistical matching will get "as good as people", i.e. still imperfect but, say, >95% confidence seems likely

# Giga-scale databases

- "Giga-scale database" (word I made up this morning):
  - Billions+ of records in many tables
  - Data gathered by multiple entities
- Errors/nulls/imperfect matches inevitable
- What are imperfect matches good for?
  - Targeted advertising (exact matches don't matter)
  - Blackmail? (public opinion does not require proof)
  - Prompt for further investigation
  - *Not* (directly) legal proceedings?
- Costs on the order of tens of millions of $ per year to "mine" this scale of data
  - Will come down with time

## Plan

- What do
  - my computer & local network
  - my Internet connection
  - websites I visit

  know about me?
- How can this data be aggregated?
- **How can we mitigate the risk of exposing "too much information"?**

## Mitigating privacy risk: technological measures

- "Separation of powers"
- Encryption
- Anonymizers
- *Post hoc* data scrubbing

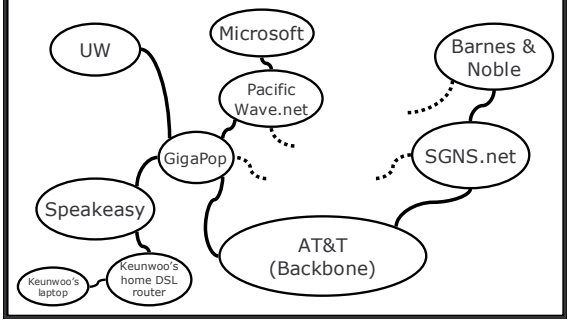
## "Separation of powers"

- Don't get your Internet connection from the company that runs your web apps
  - Depending on data you want to remain correlated, may not be effective
  - Once you buy something from Amazon, they have your IP and your name/address
- Don't get all your web services from one place
  - May reduce risk of database fusion

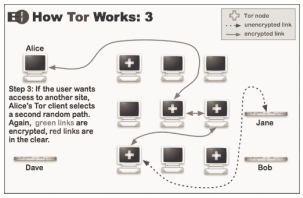
## Encryption

- Prevents *interception* of communications by third parties
- Good to have, but not the real privacy problem

## The Internet




## Anonymizers




- Severe performance penalty, probably for the foreseeable future
- Hard for novice users to set up (& who has the business incentive to make it easy?)

## *Post hoc* data scrubbing

- Why doesn't Google "scrub" its logs?
  - It wants to mine statistics
  - Hard (or sometimes impossible) to scrub data well without losing statistical properties
  - Hard to scrub data "enough" to prevent recovery by data fusion later