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)

  Keunwoo’s laptop
  Cybercafe access point

  00:0E:35:52:88:32

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

UW
Microsoft
Barnes & Noble
Speakeasy

"The Internet(s)"
The Internet

- UW
- Microsoft
- Pacific Wave.net
- Barnes & Noble

Speakeasy

AT&T (Backbone)

Internet Protocol (IP) subnets

- UW
- Microsoft
- Pacific Wave.net
- Barnes & Noble

- GigaPop
- SGNS.net

AT&T (Backbone)

IP address assignment

- Keunwoo's UW-CSE workstation: 128.208.3.7
- Keunwoo's home DSL connection: 216.254.13.6

IP sharing via NAT

- Keunwoo's UW-CSE workstation: 128.208.3.7
- Keunwoo's home DSL router: 216.254.13.67

When Google's your ISP...

- Google Web Search Server
- Google server farm
- Google municipal access point
- Google ISP

- Keunwoo's laptop: 262.64.21.71
- Neighbor's laptop: 262.64.33.11

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
  - plausible deniability?
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

<table>
<thead>
<tr>
<th>ISPCustomers</th>
<th>Telemarketinglist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Birth</td>
</tr>
<tr>
<td>Alice Acker</td>
<td>2/17/1950</td>
</tr>
<tr>
<td>Bob Booth</td>
<td>1/2/1960</td>
</tr>
<tr>
<td>Carol Collins</td>
<td>NULL</td>
</tr>
<tr>
<td>Dave Dawkins</td>
<td>3/4/1980</td>
</tr>
<tr>
<td>Dave Dawkins</td>
<td>3/6/1990</td>
</tr>
</tbody>
</table>

Databases

<table>
<thead>
<tr>
<th>columns</th>
<th>table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Birth</td>
</tr>
<tr>
<td>Alice Acker</td>
<td>2/17/1950</td>
</tr>
<tr>
<td>Bob Booth</td>
<td>1/2/1960</td>
</tr>
<tr>
<td>Carol Collins</td>
<td>NULL</td>
</tr>
<tr>
<td>Dave Dawkins</td>
<td>3/4/1980</td>
</tr>
<tr>
<td>Dave Dawkins</td>
<td>3/6/1990</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>ISPCustomers</th>
<th>Telemarketinglist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Birth</td>
</tr>
<tr>
<td>Alice Acker</td>
<td>2/17/1950</td>
</tr>
<tr>
<td>Bob Booth</td>
<td>1/2/1960</td>
</tr>
<tr>
<td>Carol Collins</td>
<td>NULL</td>
</tr>
<tr>
<td>Dave Dawkins</td>
<td>3/4/1980</td>
</tr>
<tr>
<td>Dave Dawkins</td>
<td>3/6/1990</td>
</tr>
</tbody>
</table>

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)
  - Bladorna? (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

Join on Name, selecting Name, IPAddress, Address, Merchants

<table>
<thead>
<tr>
<th>Name</th>
<th>IPAddress</th>
<th>Address</th>
<th>Merchants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice Acker</td>
<td>1.2.3.6</td>
<td>1 First St.,...</td>
<td>QFC,GAP,...</td>
</tr>
<tr>
<td>Bob Booth</td>
<td>1.2.3.7</td>
<td>3 Third St.,...</td>
<td>PornGalore,...</td>
</tr>
<tr>
<td>Carol Collins</td>
<td>NULL</td>
<td>NULL</td>
<td>Google,...</td>
</tr>
<tr>
<td>Dave Dawkins</td>
<td>1.2.3.8</td>
<td>2 First St.,...</td>
<td>PornGalore,...</td>
</tr>
<tr>
<td>Dave Dawkins</td>
<td>1.2.3.9</td>
<td>3 Third St.,...</td>
<td>Google,...</td>
</tr>
</tbody>
</table>
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

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