Surveillance technology

How use and users of digital information are tracked

Logging and logs

- Computers keep track of what happens on and to them.
  - Access logs
  - Error logs
  - Action logs (less common, because logs take up space!)
- Including web servers!
- When you affect another computer on the network, IT LOGS YOU.
  - You don't have to actually "log in" to it! Though it helps.
  - Your computer may be identified by IP address or MAC address. If you logged in, you can be identified by username.

What you can find out about a website user, (more or less) passively

- Limited technical and geographic info on site users
  - E.g. OS, browser, mobile or not, IP address (geolocatable)
- What site users do (and don't do) on your site
- What site-search-box users search for on your site
- What search terms bring people to your site from external search engines
- What links from other sites bring people to your site
  - Including social media sites, often

The old way: server log analysis

- Web-server software, like most software, keeps records ("logs") of how it is used.
- These logs are plain-text files that live in a specific (software-dependent) place on the server.
  - Each line of the log represents one "hit:" attempted file access
  - Note well: loading a whole page means a LOT of hits! One hit per HTML file, CSS file, image, external script file...
  - Browser caching can mess up the numbers too.
- These logs can be analyzed! Let's look at them.

Ack, whaaaaaaaaat?!

```
54.149.222.78 - - [31/May/2016:14:14:58 -0400] "POST /xmlrpc.php HTTP/1.1" 403 15 "-"
"Opera/9.80 (Windows NT 6.0) Presto/2.12.388 Version/12.14"
208.115.111.69 - - [31/May/2016:14:23:47 -0400] "GET /robots.txt HTTP/1.1" 404 - "-"
"Mozilla/5.0 (compatible; DotBot/1.1; http://www.opensiteexplorer.org/dotbot,
help@moz.com)"
66.249.69.174 - - [31/May/2016:14:32:22 -0400] "GET
/index.php/2015/11/04/how-to-digitize-a-vinyl-record/ HTTP/1.1" 200 7255 "-" "Mozilla/5.0
(compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
54.149.222.78 - - [31/May/2016:14:38:27 -0400] "POST /xmlrpc.php HTTP/1.1" 403 15 "-"
"Opera/9.80 (Windows NT 6.0) Presto/2.12.388 Version/12.14"
66.249.69.184 - - [31/May/2016:14:41:21 -0400] "GET /index.php/category/how-tos/ HTTP/1.1"
200 15088 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
68.180.230.228 - - [31/May/2016:14:58:01 -0400] "GET
/index.php/2015/09/02/key-tool-label-maker/ HTTP/1.1" 200 5595 "-" "Mozilla/5.0
(compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)"
54.149.222.78 - - [31/May/2016:15:01:50 -0400] "POST /xmlrpc.php HTTP/1.1" 403 15 "-"
"Opera/9.80 (Windows NT 6.0) Presto/2.12.388 Version/12.14"
151.80.31.179 - - [31/May/2016:15:11:05 -0400] "GET /robots.txt HTTP/1.1" 404 - "-"
"Mozilla/5.0 (compatible; AhrefsBot/5.1; +http://ahrefs.com/robot/)"
54.149.222.78 - - [31/May/2016:15:24:45 -0400] "POST /xmlrpc.php HTTP/1.1" 403 15 "-"
"Opera/9.80 (Windows NT 6.0) Presto/2.12.388 Version/12.14"
66.249.83.110 - - [31/May/2016:15:34:56 -0400] "GET / HTTP/1.1" 200 14681 "-" "Mozilla/5.0
(X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/49.0.2623.75
Safari/537.36 Google Favicon"
```

Looking at one line

```
54.149.222.78 - - [31/May/2016:14:14:58 -0400] "POST /xmlrpc.php HTTP/1.1" 403 15 "-"
"Opera/9.80 (Windows NT 6.0) Presto/2.12.388 Version/12.14"
```

- First bubble: User's reported IP address. Second bubble: Date/time of access
- Third bubble: What site URL the user tried to get at, and how
  - GET: get a file from the server (e.g. an HTML file or an image or whatever)
  - POST: put new information on the server (e.g. through a web form or script)
  - This user was trying to use a PHP script to put information on the server.
- Fourth bubble: HTTP response code
  - 403 means "Forbidden"! So this user was NOT authorized to be messing with this PHP file, and the web server told them so!
- Fifth bubble: the "user agent" (browser or other tool)
  - Beware: it's easy to lie about this, and many actors, nefarious and not, do.

One more… "Referer"

```
66.168.55.194 - - [01/Jun/2016:02:54:47 -0400] "GET
/wp-content/uploads/2016/03/betamax-video-switch-262x300.png HTTP/1.1" 304 -
"http://radd.dsalo.info/" "Mozilla/5.0 (iPad; CPU OS 9_3_2 like Mac OS X)
```

- This is a request for an image (".png" extension).
- "http://radd.dsalo.info/" is the "**referer**" (yes, this is an ancient spelling error), **the page that sent the user's browser to this file**.
- Time was, you could find out who was linking to your site this way. But "referer spam" and user tracking became such annoyances that browsers, websites, and even search engines may now suppress referring-page information.

Breaking out of a single-site silo

- As you know, a web page includes lots of different files: HTML, CSS, images, etc.
- It's perfectly possible (the norm, in fact) for some files on a given page to come from a different Internet domain altogether!
- This allows logging to happen (and log data to be collected) across many websites.
- In fact, **most websites add things** to their pages **just to log you**.
  - Call these "**web bugs**" or "**trackers**."

Examples of web bugs

- **Cookies**
  - ... useful for more than tracking, but OFTEN used for tracking also
  - **Third-party cookie**: Cookie set by some website that IS NOT the website you're on. You can usually disable these in your browser, and it's a good idea to do so.
- Social media **"like" and "share" buttons**
  - These log you **even if you don't click** on them!
- Tracking **scripts**, often from companies whose whole business is web tracking
- Website owners/creators know all this. They put the web bugs on their own websites!
  - So let's not pretend site owners are innocent bystanders in all this.
  - THINK HARD ABOUT WEBSITES YOU WORK ON, okay? **YOU ARE ETHICALLY RESPONSIBLE** for trackers you add.

Those ads that follow you around the web…

- Put the **same tracking bug** on **more than one website**.
- **Each website** with that bug knows about you, and **what you clicked on elsewhere**.
  - This is the usual case! It's exactly the business of Google-owned DoubleClick.
- Once the network of bugged sites decides what you'll click on, **that ad can show up on any other site** in the network.

The typical ad from a typical ad network…

- **Notices your visit** to the page via web bug
- Grabs whatever **identifying information** it can about you **from your browser**
  - cookies, browser fingerprint, referrer, etc.
- "**Phones home**" to the ad network **to learn all it can about you**
  - And **stores the data it just learned** about you with the ad network!
- Gets back the ad(s) or other content that the network thinks you're most likely to click on
  - This is **personalization**! Still think it's for your benefit?

Browser fingerprinting

- Website loggers can figure out some things about how your browser and computer are **configured**, such as:
  - which browser it is on which operating system
  - browser plugins/extensions installed and/or enabled
  - fonts installed on your computer
  - websites you're logged into, sometimes
  - websites you've visited recently, sometimes
- For most browser/device combinations, adding up all this information serves as a trackable unique identifier… a hard one to change/disguise.
  - I've seen some browser plugins that try to disguise this. I don't know how well they work.
- 2016: Firefox removes Battery Status API because of its use in fingerprinting

Tracking mobile users

- Your cell company tracks who you called, when, and for how long.
  - Similarly with text messages
- It's easy to make a fake cell tower that will fool your phone into communicating its location.
  - Such a fake is often called a "**stingray**."
- Any device seeking a wifi connection broadcasts a list of all the wifi networks it knows about.

Commercial surveillance of digital information

LIS 351

20th-century commercial surveillance

- It existed! For example, store loyalty cards (and credit cards), marketing demographic exercises, etc.
- It was sometimes opt-in with an explicit reward (loyalty cards).
- Much of it wasn't, e.g. credit scoring, but generally it was possible to find out what They knew about you.
  - This was largely due to pro-consumer regulation; the credit bureaux don't offer free credit reports out of the goodness of their hearts.

What's different now?

- Why bother with broad-brush demographic data when you can just snoop on individuals?
  - It's easy! You just need a little web bug!
- Consent? Transparency? What are those?
- Data resale and transfer much more common.
  - The "data broker" industry (yes, an entire industry!) swaps personal data around constantly... and pretty much invisibly to those tracked (that is, everyone).
  - Web tracking feeds a lot of data about individuals to data brokers!
- Pro-consumer regulation has not caught up… partly because it's legitimately a lot harder to understand what's going on!

Social media and tracking

- **User tracking** is the dominant business model for social media. It's **how social media make money**!
  - Some track mostly internally to target advertising and create other "services" (e.g. redlining!) atop the data: Facebook
  - Some sell data directly to data brokers and advertisers.
- They data-mine what you post, correlate you with those you interact with ("network analysis"), and more!
- If you're not the customer, you're the product.

So what?

- Many already-vulnerable people **become even more vulnerable** if their network activity can be tracked and/or geolocated.
  - Whistleblowers
  - Journalists, journalist informants
  - Oppressed people (who come in many, many varieties, sadly)
  - People who are being harassed or stalked or threatened/hunted, online and/or offline
  - People engaged in disapproved political activity, even when it's legal
  - Children
- And fundamentally, a lot of what we do online shouldn't be anybody else's business!
  - US law does not help here. (Eurolaw is a bit better.)

The web-tracking industry

- Purveyors of web bugs and sleazy recommenders to many websites near you!
  - **Web bug**: tiny bit of HTML or JavaScript code, or an image, placed on a web page specifically to track page visitors and their behavior.
- Often drape themselves in "**personalized recommendations**."
  - This is always, always, always a synonym for user surveillance. We talked about how recommender systems work!
  - When you can opt out of personalization, I recommend doing so.
- An adjunct to the web-advertising industry.
  - So let's talk briefly about how web advertising currently works...

Modes of web advertising

- "Native" advertising: infomercials or paid sponsorships, site-specific ads
- Small-scale advertising networks
  - These usually come together around a very specific niche market, e.g. "The Deck" aimed at web professionals.
- Native and small-scale ads don't gain much from tracking you…
  - … which doesn't mean they always avoid it, mind you.
- Giant advertising networks covering hundreds, thousands, millions of sites
  - This is most often where things get sleazy.

Targeted Advertising Considered Harmful

Targeted ad proponents tell us that ads are getting more personalized and relevant, so why isn't blocking going down instead? Why aren't users saying, "There's a magic machine in a data center that will only show me ads for stuff I really want to buy? Better turn off the ad blocker!" In another survey, 66% of adult Americans said they "do not want marketers to tailor advertisements to their interests", and when the researchers explained how ad targeting works, the percentage went *up*.

Here's the thing about that

- The process eats HUGE amounts of **bandwidth** and **page-loading time**. It's terrible, terrible usability/UX.
- The Great Ad-Blocking Controversy of 2015 was partly about this problem!
  - Especially as it intersects with mobile, where people are on expensive and limited data plans.
- The problem is especially acute on **news-media websites**.
  - These tend to have many times the number of ad networks, web bugs, recommenders that other kinds of sites do.

A short word about the "Internet of Things"

- It's really the "Internet of Things That Surveil You and Your Housemates."
  - "With Unbelievably Inadequate Security Protections"
- Most of the gadgets currently part of the IoT aren't very powerful on their own.
- To do what they do, they have to phone home to a server belonging to the gadget maker.
  - This is triply true of anything you talk to, e.g. Siri, Hello Barbie, Amazon Echo.
- Nothing says they have to use what they learn just to make the gadget work.

Thank you!

Educational surveillance

also known as “learning analytics”

Did anybody ever tell you
"this is going on your
permanent record?"

They were bluffing, mostly.

But now they're not.

FERPA

- Family Educational Rights and Privacy Act
- Protects any US-based "educational record" you have. Not just anybody can waltz in and ask to see it; usually you (if you are adult) or your parents/guardians (if not) have to consent first.
  - Even your teachers/advisors/counselors etc. have to have a reason to look up your records.
- Not perfect law, but not bad either.

Here's the thing…

- The current definition of what counts as an "educational record" is pretty specific and completely print-based.
- Lots of things it's possible to watch you do in a digital classroom don't count under FERPA.
- Add that to the "big data" movement, and you get…

"Learning analytics"

- Surveilling students as they learn, both online and in the physical world, and (supposedly) trying to use the information to help them learn.
- **Not really tested**. A lot of the "innovation" in this space is going on hunches and guesses, even as it's affecting real students.
  - I repeat: **WE DON'T EVEN KNOW IF/WHEN THIS WORKS**.
  - Worse, a lot of the experimentation is not undergoing regular research oversight.
- Obviously this information is gold to lots of other parties too…
  - … imagine if a prospective employer got hold of it. (Some of them are sleazy enough with the educational records they CAN get.)

Including…

- Anything you do in Learn@UW or similar
- E-resources you use from the library
- Website-based interactions and use, sometimes
- Anything you do in the Student Center
  - enrollment, classes you look at (without enrolling in them), etc.
- And more.
  - Higher-education institutions building "data warehouses" to retain all this information and connect it up with other information… just like a commercial data broker, really.

Wait, the physical world too? How does that work?

- ID-card swiping
  - Any time you swipe your ID, that turns into a row in a campus database, tied directly to you.
  - That includes purchases, if you use WISCARD!
- Same space-surveillance techniques that retail and law enforcement use
  - E.g. stingrays to track student cell phones

inBloom: FERPA? What FERPA?

- Gates Foundation funded effort to pool K-12 student records into one huge database
  - ... which educational-technology companies would have access to
  - ... and who knows who else?
- Did a couple of deals with large US states
- Parents and teachers HOWLED until inBloom had to admit defeat and fold.
- Lesson: it IS possible to defeat this stuff!

Issues now

- Data from educational and non-education-specific cloud apps
  - 2015: the EFF announced it's taking Google to court over this.
  - A lot of schools are diving merrily into cloud apps without thinking NEARLY hard enough about student privacy. If you have K-12 people in your family, please be aware of this and do your best to protect them.
- School hardware that tracks students
  - e.g. ChromeBooks and iPads, or anything with a remote-controllable camera
- Are learning analytics worth the privacy tradeoffs? Whom do they help? Whom do they harm, especially via data retention?
- I expect this to be an ACTIVE area of discourse and legislation in the next few years.

What students can do

- When your instructor says "Let's use this online thing!" ask back "What'll it do with my personal data?"
  - They probably won't have thought about it. At least ask them to THINK.
- Raise this with student organizations and *PIRGs.
- Use UW-Madison branded cloud apps whenever you can.
  - You're getting a better legal deal than regular consumer apps give you. Even when it's the same apps!
- Good librarians are your allies. Ask them to back you up.

Thank you!