Privacy

CS 161 - Computer Security
Profs. Vern Paxson & David Wagner

TAs: John Bethencourt, Erika Chin, Matthew Finifter, Cynthia Sturton, Joel Weinberger
http://inst.eecs.berkeley.edu/~cs161/

March 31, 2010
Announcements

• Reminder: on Friday go to 1 Pimental, not here, for Midterm #2
  – 5:10-6:30PM
  – You can bring a single page “cheat sheet”
    • Plus you can also bring the cheat-sheet from Midterm #1

• Note: no section next week
Defining Privacy

• Privacy = right to control who knows certain aspects about you / your communications / your activities
  – Control over disclosure
  – And ideally over subsequent use

• How much of an issue is this?
  E.g., how much information about you do web sites learn as you surf?
Privacy & Web Surfing

• The sites you visit learn:
  – The URLs you’re interested in
    • Google/Bing also learns what you’re searching for
  – Your IP address
    • Thus, your service provider & geo-location
    • Can often link you to other activity including at other sites
  – Your browser’s capabilities, which OS you run, which language you prefer
  – Which URL you looked at that took you there
    • Via “Referer” header
Privacy & Web Surfing, con’t

• Oh and also cookies.
• Cookies = **state** that server tells browser to store locally
  – Name/value pair, plus expiration date
• Browser returns the state any time visiting the same site

• Where’s the harm in that? And are these used much anyway?
Let's remove all of our cookies.
We do a Google search on “private browsing” and click on the top result.
Note that this mode is privacy from **your family**, not from web sites!
Whoa - we gained 11 cookies!

What on earth is Google tracking in this one?

It sticks around for 6 months.
Hmmm. Mozilla is tracking us too. And for 5 years!
They’re even remembering just how we visited them.
And something else (as we’ll see in a bit) until the End Of Time
Without doing anything else, we’ve gained a 12th cookie …
We now do just one more operation, opening the home page of www.nytimes.com
doubleclick.net - who’s that?  
And how did it get there from visiting www.nytimes.com?  
What a lot of yummy cookies!
Third-Party Cookies

• How can a web site enable a third party to plant cookies in your browser & later retrieve them?
  – Answer: using a “web bug”
  – Include on the site’s page (for example):
    • `<img src="http://doubleclick.net/ad.gif" width=1 height=1>`

• Why would a site do that?
  – Site has a business relationship w/ DoubleClick*
  – Now DoubleClick sees all of your activity that involves their web sites (each of them includes the web bug)
    • Because your browser dutifully sends them their cookies for any web page that has that web bug
    • Identifier in cookie ties together activity as = YOU

* Owned by Google, by the way
Remember this till-the-End-of-Time cookie?
Google Analytics

• Any web site can (anonymously) register with Google to instrument their site for analytics
  – Gather information about who visits, what they do when they visit
• To do so, site adds a small Javascript snippet that loads http://www.google-analytics.com/ga.js
  – You can see sites that do this because they introduce a "__utma" cookie
• Code ships off to Google information associated with your visit to the web site
  – Shipped by fetching a GIF w/ values encoded in URL
  – Web site can use it to analyze their ad “campaigns”
  – Not a small amount of info …
Values Reported via Google Analytics

Affiliation
Billing City
Billing Country
Billing Region
Browser Lang.
Complete URL
Cookie Values
Current Page
Event Tracking
Flash Version
Grand Total
Host Name
Java-enabled
Language Encoding
Order ID
Page Title
Product Code
Product Name
Profile Number
Repeat Campaign Visit
Quantity
Screen Color
Screen Resolution
Shipping Cost
Special Event
Start Campaign Sess.
Tax
Tracking Code Version
Unique GIF ID
Unit Price
User Defined Var
Variations on an Item
Privacy - What’s the Big Deal?

• Cookies form the core of how Internet advertising works today
  – Without them, arguably you’d have to pay for content up front a lot more
    • (and payment would mean you’d lose anonymity anyway)
  – A “better ad experience” is not necessarily bad
    • Ads that reflect your interests; not seeing repeated ads

• But: ease of gathering so much data so easily ⇒ concern of losing control how it’s used
  – Mission creep …
    • Consider how ordering a pizza in the near future might work (http://www.aclu.org/ordering-pizza)
  – Content shared with friends doesn’t just stay with friends …
More Employers Screening Candidates via Social Networking Sites

Five tips for creating a positive online image
Rosemary Haefner, Vice President of Human Resources at CareerBuilder

Gone are the days when all job seekers had to worry about were their résumés and cover letters. Today, those documents remain a staple of the job-search process, but they are joined by a growing phenomenon: social networking.

Forty-five percent of employers reported in a June 2009 CareerBuilder survey that they use social networking sites to screen potential employees, compared to only 22 percent of employers last year. Eleven percent of employers plan to start using social networking sites for the screening process. More than 2,600 hiring managers participated in the survey.
Why employers disregard candidates after screening online

Thirty-five percent of employers reported they have found content on social networking sites that caused them not to hire the candidate, including:

- Candidate posted provocative or inappropriate photographs or information -- 53 percent
- Candidate posted content about them drinking or using drugs -- 44 percent
- Candidate bad-mouthed their previous employer, co-workers or clients -- 35 percent
- Candidate showed poor communication skills -- 29 percent
- Candidate made discriminatory comments -- 26 percent
- Candidate lied about qualifications -- 24 percent
- Candidate shared confidential information from previous employer -- 20 percent
How To Gain Better Privacy?

• Force of law
  – Example #1: web site privacy policies
    • US sites that violate them commit false advertising
    • But: policy might be “Yep, we sell everything about you, Ha Ha!”
  – Example #2: SB 1386
    • Requires an agency, person or business that conducts business in California and owns or licenses computerized 'personal information' to disclose any breach of security (to any resident whose unencrypted data is believed to have been disclosed)
    • Quite effective at getting sites to pay attention to securing personal information
Gaining Privacy Through Technical Means

• How can we surf the web truly anonymously?
• Step #1: remove browser leaks
  – Delete cookies (oops - also “Flash cookies”!)
  – Turn off Javascript (so Google Analytics doesn’t track you)
• Step #2: how do we hide our IP address?
• One approach: trusted third party
  – E.g. anonymizer.com
    • You set up an encrypted VPN to their site
    • All of your traffic goes via them
  – Issues?
    • Performance
    • ($80/year)
    • “rubber hose cryptanalysis” (cf. anon.penet.fi & Scientologists)
Anonymous Web Surfing, con’t

• Idea: remove single point of trust failure by chaining together a series of servers
• Suppose Alice wants to send a message \( X \) anonymously with Bob
• And there are \( N \) servers, \( M_1 \) \ldots \( M_N \) ("mixes"), available, each with a public key \( K_1 \) \ldots \( K_N \)
  – Each mix will accept a (message, next-hop) pair encrypted w/ its key and forward message to the mix (or end system) given by the next hop
• Approach: Alice bounces her message among the mixes to mask its origin ("onion routing")
Peeling the Onion

- Alice picks some mixes at random, say $M_i$, $M_h$ & $M_k$
- She sends to $M_i$ the following:
  \[ \{ \{ \{ X, B \}^{K_k}, M_k \}^{K_h}, M_h \}^{K_i} \]
- $M_i$ receives $\{ \{ \{ X, B \}^{K_k}, M_k \}^{K_h}, M_h \}^{K_i}$, decrypts
  - Message inside is $\{ \{ X, B \}^{K_k}, M_k \}^{K_h}$, next hop is $M_h$
- $M_h$ receives $\{ \{ X, B \}^{K_k}, M_k \}^{K_h}$, decrypts
  - Message inside is $\{ X, B \}^{K_k}$, next hop is $M_k$
- $M_k$ receives $\{ X, B \}^{K_k}$, decrypts
  - Message inside is $X$, next hop is $B$
- B receives $X$; has no idea who sent, nor does $M_h$/$M_k$
- Note: this is what the industrial-strength Tor anonymizing service uses
  - It also provides bidirectional communication
Onion Routing Issues/Attacks?

- Performance: message bounces around a lot
- Key management: the usual headaches
- Attack: rubber-hose cryptanalysis of mix operators
  - Defense: use mix servers in different countries
    - Though this makes performance worse :-(
- Attack: adversary operates $M_i$
  - Defense: have lots of mix servers (Tor today: ~2,000)
- Attack: adversary observes when Alice sends and when Bob receives, links the two together
  - A “confirmation” attack
  - Defenses: pad messages, introduce significant delays
    - Tor does the former, but notes that it’s not enough for defense
Onion Routing Attacks, con’t

• Issue: leakage
• Suppose all of your HTTP/HTTPS traffic goes through Tor, but the rest of your traffic doesn’t
  – Because you don’t want it to suffer performance hit
• How might the operator of sensitive.com deanonymize your web session to their server?
• Answer: they inspect the logs of their DNS server to see who looked up sensitive.com just before your connection to their web server arrived
• Hard, general problem: anonymity often at risk when adversary can correlate separate sources of information
Dataset Privacy

• Difficult issues of anonymity arise when releasing database records

• Recent example: Netflix released a portion of their customer records in a contest to improve their recommendation system
  – Data included anonymized user ID, some of the movies user rated, how much the user liked them, and when user rated them

• How could (some) users be deanonymized?

• Attackers (researchers) cross-correlated with non-anonymous IMDB movie reviews
  – Looked for rarely-reviewed movies for which same movie was reviewed in Netflix & IMDB at about the same time

• General finding: in datasets with modest level of details, individuals tend to be in some way unique

• Related finding: birthdate + gender + zip code = unique for 60+% of US population! (note, P&P quotes older 87% figure)
My browser had Flash cookies from 67 sites!

Sure, this is where you’d think to look to analyze what Flash cookies are stored on your machine.

Some Flash cookies “respawn” regular browser cookies that you previously deleted!

My browser had Flash cookies from 67 sites!
7. Collection of Viewing Information. You acknowledge that you are aware of and consent to the collection of your viewing information during your use of the Software and/or Content. Viewing information may include, without limitation, the time spent viewing specific pages, the order in which pages are viewed, the time of day pages are accessed, IP address and user ID. This viewing information may be linked to personally identifiable information, such as name or address and shared with third parties.