Lecture 3: Technology – 1

Balachander Krishnamurthy

AT&T Labs–Research

`http://www.research.att.com/~bala/papers`

Papers to read this week

[KW06] Generating a Privacy Footprint on the Internet

[KW09] Privacy Diffusion on the Web: A Longitudinal Perspective

[MM12] Third-Party Web Tracking: Policy and Technology

[Mar10] Abusing social networks for automated user profiling

[GPS09] KnowPrivacy: The Current State of Web Privacy, Data Collection and Information Sharing

[KNW11] Privacy leakage vs. Protection measures: the growing disconnect

Problems with k-anonymity

- Recall 87% of Americans being uniquely identified via zipcode, gender, and birth date: these identifiers are called quasi-identifiers
- In the released data, quasi-identifier must be present in at least k-records
- k-anonymity problems: homogeneity and background knowledge attacks
- Homegeneity: Neighbour may have partial information (e.g., same zip code, rough age) and can narrow k. So if n people have cancer in that zip code then neighbour having cancer can be deduced
- Background knowledge: If certain nationalities have low incidence of a particular disease, that can be used to reduce k and potentially identify someone

l-diversity addresses these problems: [MKGV07]

Lecture 3: Technology–1

- **Terminology and key players**
- Tracking
- Technologies for tracking
- Identifying leakage
- Role of JavaScript
- Role of protocols

Terminology

- First party: user sets up direct communication by clicking on a link or entering URL

- Third party: browser auto-redirected to such sites

  Could be outsourced site (CDN, analytics)
  Aggregator/advertiser

- Leakage: Information sent to a party without informed consent of user

- Linkage: Merging information across different sites and services

- Behavioral tracking: typically long term gathering of user browsing information

Who are the key players?

- Users
- Publishers
- Aggregators and third parties in general
- Moderators
- Large, somewhat visible commercial entites
- Larger hidden ecosystem

Moderators

Privacy organizations

- Privacy International (UK, ’90, 46 countries), EPIC, CDT (offshoot of EFF ’90)
- Several more...
- Activists, privacy advocates, researchers

Governmental agencies:

- FTC (US)
- Provincial Privacy Commissioner (Canada),
- European Data Protection Supervisor

Large somewhat visible commercial entites

- IAB–Interactive Advertising Bureau (500 cos., 86% of online ads)
- MMA–Mobile Marketing Association (700 cos.)
- Data exchange

  BlueKai (audience stitching)
  Rapleaf (1B email)
  Acxiom (customer information infrastructure)

Tracking

- No accepted definition of tracking yet!
- EFF says: “Tracking is the retention of information that can be used to connect records of a person’s actions or reading habits across space, cyberspace, or time” https://www.eff.org/deeplinks/2011/02/what-does-track-do-not-track-mean
- CDT says “Tracking is the collection and correlation of data about the Internet activities of a particular user, computer, or device, over time and across non-commonly branded websites, for any purpose other than fraud prevention or compliance with law enforcement requests” https://cdt.org/blogs/erica-newland/cdt-releases-draft-definition-”do-not-track”

Views on tracking

- Shadowing of users' movements on the Internet can be a loose definition
- Somewhat creepy depending on point of view
- Tracking can be done by first party, via outsourced analytics, or via third parties
- Note that data retention is often mandated by law!
- Advertisers: We want to provide *targeted* advertising and thus knowing user's movements let us infer interests
- Aggregators: we help advertisers and first party sites at their request

Reasons to track

- Site loading evaluation (improve performance)
- Simpler site navigation (no need to re-enter passwords etc.)
- Enhancing user experience (typical use of JavaScript)
- Learning demographics of site (re-orient content)
- User behavior study (effective positioning of content)
- Results of reconfiguring site (improving site)
- Targeted advertising (monetization)

Technologies for tracking

Several broad categories

1. Cookies (still evolving as recently as this past week..)
2. Embedding links in Web pages
3. Potentially via outsourcing to CDNs
4. JavaScript

1. Cookies

- HTTP is stateless: Web servers do not have to retain information about past requests
- But this might be needed for facilitating return visits by same user
- State management is provided via opaque strings called *cookies* (see RFC 6265)
- Cookies are a two-decade old innovation and still in wide use
- Executive summary: service sends a Set-Cookie response header with the cookie, clients then send back the cookie in the Cookie request header
- Cookies have lifetimes associated with them (session-specific, years)
- For more details on cookies See Chapter 2 of [KR01]

[KR01] Web Protocols and Practice: HTTP/1.1, Networking Protocols, Caching, and Traffic Measurement

Potential uses of cookies

- Simple way to correlate users across Web sessions...
- ...without maintaining information on server end for millions of users
- Simplifies shopping cart applications so users do not have enter identifying information each time

Cookies: user control

- Users can disallow setting of cookies
- Allow only for current session
- Limit origination of cookies to first party site
- Delete cookies at any time
- Rarely done by vast majority

Known privacy problems with cookies

- Given that they are opaque strings, exact information sent via cookies is unknown

- Links in hidden back-end database by servers can make cookies persist beyond user's expectation (re-identification and re-linking possible)

- Third-party servers sending cookies can be problematic (we will see a detailed example of this issue later)

- Different 3rd-parties could share cookie information and correlate them to construct a broader user profile

- In spite of cookies origination in 1994, there is little that is understood about their use by vast majority of users

2. Embedding links in Web pages

- Since the creation of 3rd-parties, the easiest way is to embed links that are auto-download

- 3rd-parties work in conjunction with interested first parties who must see value in embedding links to them

- First parties get potentially valuable information from such embedding

- The same 3rd-parties are present in multiple first party Websites

- Users can see the additional 3d-party interactions but no easy interactive way to block (too many)

- (Later we will look at automated techniques to block such interactions)

3. Potentially via outsourcing to CDNs

- CDN: Content Distribution Networks
- E.g. Akamai, Limelight, Level3
- Saves server load on first parties, improve delivery speed
- CDNs may be interested in the data they get from being present on multiple first party sites

4. JavaScript

- Downloaded and interpreted in the browser
- Wide variety of scripts; most used to improve site experience
- Indispensable in maps and many other applications
- Also used in tracking
- Code interpreted in browser's memory and thus has access to state
- Can deposit output in cookies or other HTTP headers and send back to server

Identifying leakage

- Earlier you saw examples of 'hidden' sites visited as a result of visiting first party sites
- Later I will describe a 6-year long *footprint* study of tracking the trackers
- First, we will look at *techniques* by which we can identify leakages
- We begin by defining *leakage*: depends on viewpoint!

  User: Personal information shared with any site other than first party
  First party: We outsource work to third party (e.g. for analytics).
  Tracking by third party for marketing/demographic information may also be leakage.

Third parties

- Ad Networks: First-party sites (publishers) arrange with ad networks to place ads on their pages via images or javascript code.
  E.g., Google's Adsense (googlesyndication.com, doubleclick.net),
  AOL (advertising.com, tacoda.net), Yahoo!(yieldmanager.net)

- Analytics companies: measure traffic, characterize users by downloading a JavaScript file and send back information in a URL.
  E.g., google-analytics.com (urchin.js), 2o7.net (Omniture),
  atdmt.com (Microsoft/aquantive), quantserve.com (Quantcast)

- CDNs: Serve images, rarely JavaScript. e.g., akamai.net, yimg.com

Privacy could leak to all of them.

Footprint study

- Examine the number and diversity of 3d-party sites visited as a result of a user visiting first party sites.
- Look at the 3d-party domains aggregating information over time (N.B. multiple 3d-parties may track users on a single first-party site)
- Visible nodes: Popular 1200 Web sites in dozen Alexa categories
- Extracted hidden nodes corresponding to each visible node via a Firefox extension that fetches objects and records request/response
- Examined cookies, JavaScript, identifying URLs (those with ? = &)
- Also narrowed examination to *consumer* and *fiduciary* sites: subset of sites that raise more privacy concerns.
- Study carried out roughly twice a year since October 2005

Categories of 3d-party domains

1. Only set 3d-party cookies, no JS (dclk, atdmt, 2o7.net)
2. Use JS with state saved in 1st-party cookies (google-analytics: urchin.js examines 1st-party cookies, forces retrieval via an identifying URL to send information to 3d-party server)
3. Both 3d-party cookies and JS to set 1st-party cookies (quantserve)
4. 3d-party cookies and JS not used to set 1st-party cookies but serve ad URLs with tracking information (adbrite, adbureau)

Role of protocols in tracking

- Multiple protocols are involved in a typical Web transaction
- Protocols are opaque to virtually all end users
- Several attempts have successfully been made to exploit tracking via "clever" uses of different protocols
- Application level leakages are difficult to locate; identifying leakages via protocol-based techniques significantly harder
- Unlike embedded links in HTML (which are visible, hard to change quickly) external protocol-related databases can be modified
- Guarantees of full breadth examination harder
- Unusual interactions between protocols and other tracking infrastructure (e.g. Cookies)

DNS role in tracking

- Introduces a necessary degree of opacity
- DNS infrastructure plays a role
- Notion of ADNS: Authoritative DNS server
- Responsible for resolving queries related to domains
- Websites can and do outsource this
- Sub-domains can be made to appear similar at the surface level
- Who is responsible to resolve metrics.cnn.com?
- What does it resolve to?

References

[GPS09] Joshua Gomez, Travis Pinnick, and Ashkan Soltani. Knowprivacy: The current state of web privacy, data collection and information sharing, June 2009. http://knowprivacy.org/report/KnowPrivacy_Final_Report.pdf.

[KNW11] B. Krishnamurthy, K. Naryshkin, and C. Wills. Privacy leakage vs. protection measures: the growing disconnect. In *Web 2.0 Workshop on Security and Privacy*, May 2011. http://www.research.att.com/~bala/papers/w2sp11.pdf.

[KR01] Balachander Krishnamurthy and Jennifer Rexford. *Web Protocols and Practice: HTTP/1.1, Networking Protocols, Caching, and Traffic Measurement*. Addison-Wesley, May 2001. ISBN 0-201-710889-0.

[KW06] Balachander Krishnamurthy and Craig E. Wills. Generating a privacy footprint on the Internet. In *Proceedings of IMC*, October 2006. http://www.research.att.com/~bala/papers/pfp-imc06.pdf.

[KW09] Balachander Krishnamurthy and Craig E. Wills. Privacy diffusion on the web: A longitudinal perspective. In *WWW*, 2009. `http://www.research.att.com/~bala/papers/www09.pdf`.

[Mar10] Marco Balduzzi, et al. Abusing social networks for automated user profiling. In *RAID*, 2010. http://www.iseclab.org/papers/raid2010.pdf.

[MKGV07] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthu Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Transaction on Knowledge Discovery from Data (TKDD)*, 1(3), 2007.

[MM12] Jonathan R. Mayer and John C. Mitchell. Third-party web tracking: Policy and technology. In *Proceedings of IEEE Symposium on Security and Privacy*, May 2012. `https://stanford.edu/~jmayer/papers/trackingsurvey12.pdf`.