Lecture 7: Protection–1

Balachander Krishnamurthy

AT&T Labs–Research

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

Reminder: Homework assignment 3 – due TODAY!

You should have received feedback by now. Revised proposal is to help you write the design document due 3/26 ($as\ always\ NO\ extensions$). In the design document clearly identify each of the team member's contributions. Ideally you will finish the design document before embarking on coding but not sure if you will have time to delay! It depends on your confidence-level and to what extent you have broken down the project into meaningful components.

Group projects vs. 1 person projects

- Group projects were less ambitious than they should be
- I don't want grade shock to hit anyone but note that you are going to write the paper together.
- Working together can reduce overall time as well.
- This is not true for those working by themselves
- So expectation of overall work will be superlinear
- You need to say exactly what you plan to do/did and others have to agree with it

Lecture 6: Technology – 4

- Linkage
- Mechanics of detecting leakage
- **Mechanics of detecting and examining linkage**
- Semantics and the compositional problem
- Collateral Damage of Privacy

Mechanics of detecting and examining linkage–1

[YXY+12] Host Fingerprinting and Tracking on the Web: Privacy and Security Implications

examines email traces, search engine traces and cross-check identifies to see how host fingerprints are identifiable.

[JJLS10] An Empirical Study of Privacy-Violating Information Flows in JavaScript Web Applications

studies 50K popular Websites to see how often cookies are stolen, locations are hijacked, history is sniffed etc. They implemented an information flow engine in the browser, and injected taints to see its spread.

Mechanics of detecting and examining linkage–2

[Eck10] How Unique Is Your Web Browser?

examined several HTTP headers (User Agent, Accept, Cookie), along with screen resolution, timezone, plugins and their vesions, fonts etc to define entropy

Asked people to visit panopticlick.eff.org (crowdsourced data collection) and endd up with nerly half a million fingerprint instances

Lecture 6: Technology – 4

- Linkage
- Mechanics of detecting leakage
- Mechanics of detecting and examining linkage
- **Semantics and the compositional problem**
- Collateral Damage of Privacy

Semantics and the compositional problem

- Lecture 1: Gathering and sharing information must be specific to that context
- Looking up information in a library is different from searching online
- Contexts should constrain how information flows (Nissenbaum)
- Composition: information gathered over time from different sources
- Semantics: examine the lifecycle of private data collected, its duration, and the context in which it ends up being used

Definition of semantics

- The study of language meaning or study of the relationship between words and meanings
- Differentiating syntax and semantics: colorless green ideas sleep furiously
- Is this sentence grammatically correct?
- Does it mean anything?

So how does this relate to privacy?

Does it make sense to talk about privacy without

- an understanding of the full flow of data
- all the parties involved
- duration of persistence of data
- manner in which it could be used

Else it suffers from the Chomskian example problem: grammatically correct but has no meaning

Composition

- Single source analysis is not useful either
- Full visibility needs to be obtained
- Information needs to be gathered over time
- Appropriate linkage potential needs to be examined
- Which combination of elements can help reidentify quickly?
- Focus on the key combinations as a way of prevention
- Likewise, among the various semantic components: longitudinal axis, set of entities that can obtain data, manner of use etc., see which ones can contribute more to the privacy problem
- To get maximum payoff for leakage prevention focus on higher payoff components

Protection approaches

In lectures 7 and 8 we will examine different protection techniques

Note that your project can be future protection measures!

Lecture 6: Technology – 4

- Linkage
- Mechanics of detecting leakage
- Mechanics of detecting and examining linkage
- Semantics and the compositional problem
- **Collateral Damage of Privacy**

Collateral Damage of Privacy

- Severely understudied problem
- Real world understanding is high
- If I leak information about another person that is collateral damage
- Often such leakage may happen accidentally (at times it may be deliberate)
- Detection is hard (reputation.com)
- Prevention is virtually impossible

Should we give up on CDP?

- No! In fact I think this is a low hanging fruit
- Either for a project and/or a research paper
- Understudied does not mean it is impossible at least to locate/quantify
- Almost any prevention attempt would be a contribution
- Thoughts?

Questions on last lecture?

Volunteer to summarize (5-10 min presentation) [lin] Record linkage and privacy

[RKW12] Detecting and defending against third-party tracking on the web

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

Rather than presenting papers, it'd be good if you summarize 3 privacy protection tools each, among the various tools out there.

Many of you have already presented. Some have not: you know who you are. Can you please volunteer or else I'll have to assign papers..

Lecture 7: Protection – 1

Privacy protection: a taxonomy

- What to protect?
- Who to protect against?
- Where to provide protection?
- Architectural issues
- Degree of user control
- Organizational issues and scaling up

Taxonomy: What to protect?

- Without knowing all the elements that are being gathered to violate privacy protection it is hard to know what exactly to protect
- Situation harder when ambient data is factored in
- We already know that the precise set of privacy related "bits" that are needed to re-identify have not been fully enumerated
- While gender, zip, and date of birth have been shown as quite helpful in identifying someone, we do not know all combinations

Taxonomy: Who to protect against?

- Do we worry about first parties, third parties, everyone?
- Further, we do not yet know exactly what bits of information are leaked to whom
- We do not know which external entities can merge collected information
- We do not know what public information already exists for ambient merger

Taxonomy: Where to provide protection?

- Browser
- At an intermediary (organizational proxy)
- Specific slices of the protocol stack
- Other places?

Taxonomy: Architectural issues

- Where to store user data?
- Currently virtually all OSNs store data centrally
- Distributed storage raises usual concerns regarding data synchronization
- Research projects have considered alternatives (Lockr, Vis-a-vis)

Distributed store

- Several research proposals
- All in the cloud
- Partial separation onto cellphones and some in proxies
- Clone users and use avatars for privacy
- None have solved the economic question

Taxonomy: Degree of user control

- Ideally each user will be able to tune the degree of protection
- However, as we have seen, many tend not to change defaults
- So defaults must be strong or value of protection is diminished
- Usability plays a strong role here in control (as we will see later)

Early protection techniques

- Preventing connection setups
- Identifying hidden connections and circumventing them
- Targeting large third parties
- Avoiding cookies and JavaScript
- Deleting specific headers (Referer, Cookie etc.)
- Anonymization at IP level etc.
- Opting-out (when there is a way to do that)

Early set of tools tend to follow this taxonomy

Early tools

- AdBlock: not connecting means no direct data sharing
- AdBlock Plus is now one of the most popular extensions on Firefox (19M)
- NoScript (privacy is an added value) (2.25M)
- Ghostery (set of sites to block) (1M)
- DoNotTrackMe (later addition)
- Bugmenot
- Trackmenot
- Now over a thousands privacy related extensions in Firefox alone

Closer look at AdBlock Plus

- Most common blocking technology: no connection, no data shared
- REs are relatively easy to construct given default URL patterns
- They can be shared in different groups and updated periodically
- Easy to crowdsource
- Allows whitelisting (rumours of paid whitelisting..)
- However, since it is public, advertisers can trivially modify URLs to thwart or create churn

Tool specific for OSNs?

- As can be imagined, external tools that can help with individual OSN settings is hard
- Most do not have open APIs that allow access to privacy settings
- This requires more work on the part of individual users to do work
- It would be interesting to think of ways of addressing this
- Most of the work thus far (including my research) has been more along the lines of raising awareness

Other crowdsourced techniques

- WOT: Web of Trust toolbar in IE
- Aggregates ratings of many volunteers on reliability of sites
- Simple traffic signal color scheme
- Allows customizability: Parental control (strict), light, basic etc.
- Other popular IE add-ons include Better Privacy
- Chrome has AVG PrivacyFix (you may want to look around for more)
- W3C's Privacy Dashboard http://code.w3.org/privacy-dashboard

Difficulties in blocking JS

- Semantic complexity
- RE blocking: prevention of connection set up is clear even if REs may be hard to parse
- But prevention of execution of code, when code is obscure is much harder to understand
- Automatically leads to 'trusted' blocking (which led to the AdBlock Plus whitelisting)
- No clear automatical explanation potential for random JS
- Hard to study impact of JS except on a case-by-case basis (tedious, time consuming)

NoScript interaction problems

- Note that the same piece of JS can be used to do some necessary computation in addition to (optional) tracking
- If the script is blocked, some part of the transaction won't complete
- The user may not be aware of partial execution!
- It is hard for server end to detect if the problem was due to selective blockage
- This has actually occurred on e-commerce sites
- Common solution is warn users that the site absolutely requires JS to be turned on

Countermeasures: Cat and Mouse game

- When privacy protection becomes prevalent it can impact ad revenue
- Technical papers have demonstrated this (most recently [GEC+13])
- Unsurprising that trackers and others will react
- Several simple steps have already been taken
- Blocking is relatively easy to detect (initial request is not followed by embedded requests). Some websites notify users about such blocking
- JavaScript blocking is trickier; sites can legitimately warn people of site "breakage"
- Worse yet, affect presentation significantly forcing users to allow JS

Advantage: publishers and trackers

- The edge by default will not be in user's favor
- Why?
- Publishers/trackers have longitudinal data on well-behaved clients facilitatating identification of outliers
- Publisher owns content and thus has complete control on who can access it
- Can move to subscription-based model if too many users evade tracking
- Might affect CPM metric (which is why many sites use hybrid model of paywall)
- Currently use of protection is by an insignificant minority; hence model remains open

Role of usability

- Increasing number of people are aware of privacy problems
- Reasonable number of privacy protection tools available
- While not comprehensive, reasonable degree of protection is feasible
- Other extensions are used extensively
- Yet, the fraction of users who use privacy tools is still quite low
- Usability may be a contributing cause (default settings are probably a stronger factor)

Usability improvement

- Visible feedback to users
- At all times. Or at session start/end
- Summary of leakage at end of session...
- ...with links to modify settings
- Customizable severity ratings for various settings

Cost of protection

- There are now companies (`myid.com, reputation.com`) that will help protect your privacy
- Other sites include Abine, lifelock, identityguard, allclear, trustedid etc.
- That is a direct cost when outsourced
- But there are hidden costs: what happens if you do not share allegedly required information
- Turning off images may not be terrible (captcha might fail)
- Turning off JS may break web site

References

[Eck10] Peter Eckersley. How unique is your web browser? In *Privacy Enhancing Technologies*, volume 6205, pages 1–18. 2010. https://panopticlick.eff.org/browser-uniqueness.pdf.

[GEC+13] Gill, Erramilli, Chaintreau, Krishnamurthy, Papagiannaki, and Rodriguez. Follow the money: Understanding economics of online aggregation and advertising. In *Proceedings of IMC*, October 2013.

[JJLS10] Dongseok Jang, Ranjit Jhala, Sorin Lerner, and Hovav Shacham. An emperical study of privacy-violating information flows in JavaScript web applications. In *In Proc. of ACM CCS*, October 2010. http://cseweb.ucsd.edu/~hovav/dist/history.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.

[lin] Record linkage and privacy: Issues in creating new federal research and statistical information. GAO-01-126SP Report April 2001. http://www.gao.gov/new.items/d01126sp.pdf.

[RKW12] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. Detecting and defending against third-party tracking on the web. In *Symposium on Networked Systems Design and Implementation*, April 2012. http://www.franziroesner.com/pdf/webtracking-NSDI2012.pdf.

[YXY+12] Ting-Fang Yen, Yinglian Xie, Fang Yu, Roger Peng Yu, and Martin Abadi. Host fingerprinting and tracking on the web: Privacy and security implications. In *Proceedings of the Network & Distributed System Security Symposium*, February 2012. http://research.microsoft.com/pubs/156901/ndss2012.pdf.