Lecture 4: Technology – 2

Balachander Krishnamurthy

AT&T Labs–Research

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

Papers to read this week

[KW08] Characterizing Privacy in Online Social Networks

[Kri09] Privacy Leakage in Mobile Online Social Networks

[KW09] On the Leakage of Personally Identifiable Information Via OSNs

[Mar10] Abusing social networks for automated user profiling

[KW10] Privacy Leakage in Mobile Online Social Networks

[WCG+11] Privacy Revelations for Web and Mobile Apps

Volunteer to present on SnapChat, Instagram

PII: What is personally identifiable information?

PII is defined as information which can be used to distinguish or trace an individual's identity either alone or when combined with other public information that is linkable to a specific individual.

Often the definition includes the notion of identifying in context.

Not all countries agree on a specific definition. Not all companies seem to agree on it either.

Is IP address PII?

U.S. District Court Judge Richard Jones in 2009 said, "No" and claimed it identified a computer and not a person. Case was against MSFT for auto-updating anti-piracy software and inter alia collecting IP addresses.

NIST document Special Publication 800-122

http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf

by McCallister, Grance, Scarfone gave examples:

Name, such as full name, maiden name, mothers maiden name, or alias

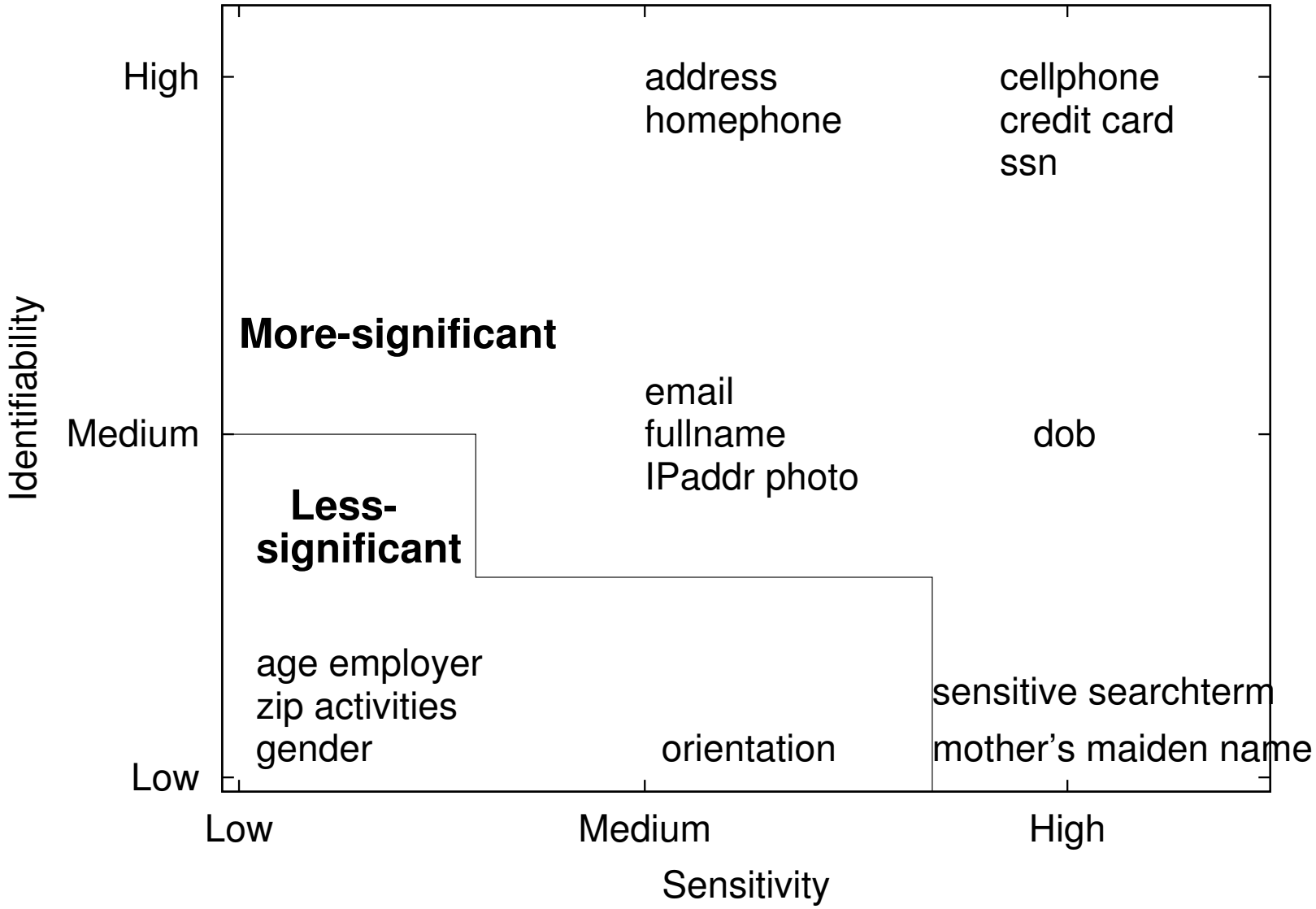
Personal identification number: SSN, passport number, drivers license number, taxpayer id, or financial account or credit card number

Address information, such as street address or email address

Personal characteristics, including photographic image (especially of face or other identifying characteristic), fingerprints, handwriting, or other biometric data (e.g., retina scan, voice signature, facial geometry)

Information that is linked or linkable to one of the above (e.g., date of birth, place of birth, race, religion, weight, activities, geographical indicators, employment information, medical information, education information, financial information).

Sensitivity and identifiability of bits of personal information



People search engines

- A lot easier to find someone than one might think...
- ...even if they are not on FB etc.
- Not many people go out of their way to hide these days
- But people search engines find people and information about people
- Large enough industry that there is competition

Examples of people search engines

I am not endorsing any of these!

- Some well-known names include: 123-people, spokeo, pipl, wink, intelius
- Some not-so-well-known names: Abika, findwhoo, Peekyou, Yasni
- http://www.findfree-people-friends.com/
- Tweepz searches in Twitter
- Many bait and switch by saying they and free and once they have gathered information they will ask for money!

Role of people search engines in privacy debate

Most PSEs provide legally available information often culled from city hall records, registration records etc.

These are ambient databases that can be used to link information.

All this information was always available except in paper form.

Now they are indexed and available online...

and so eminently linkable!

Such linkage can lead to trivial reidentification

Sweeney's paper used the information that William Weld had recently fainted in a public event and had gone to a hospital and generated a medical record.

Online Social Networks

OSNs are a subset of Web 2.0 sites, adopting most innovations with several platform/OSN-specific ones.

- Over 1 in 3 users who have access to Internet are on a OSN
- External applications going viral
- Open API radically alters picture (20x growth in Twitter)
- API allows arbitrary developers to build external applications, extend reach of OSN
- Mobile OSNs are already becoming dominant (more than 40% of all Facebook accesses are from mobile devices)
- 53% of Facebook's ad revenue comes from mobile (sent their stock up)

A simple graph model for OSNs

- Graph G with nodes V and edges E
- OSN is really an entity-interaction network which can be better modeled via multigraphs or hypergraphs
- A hypergaph H has collections of nodes (one set of nodes for each type) and edges connecting pairs of nodes
- Decision on type of node or edge or which node/edge attributes to include depends on several questions regarding node and edge properties

One reason to use a graph model is it lets us reason about OSNs through the prism of a graph.

Node properties

- How is a new node created (by user action, by site)?
- How are various node properties set (by user, site, 3rd party data, from a digital object's properties)?
- What determines the life time of a node (creation time, deletion time)?
- Where does a node "live" in the network (i.e., where can it be observed/measured)?
- How can a node be found by a user (via search, links from other objects)?

Edge properties

- What types of entities are linked by a particular edge?
- Are the edges directed or undirected? Can directed edges be reciprocated (a matching edge in the reverse direction added)?
- Who can create an edge from entity A to entity B: the owner of A? the owner of B? either? other users in the site? by the site itself?
- Whose consent is required for the creation of a proposed edge, if any? (the owner of the entity receiving the link? the site?)
- Can edges be deleted, and if so, by whom?
- What determines the lifetime/shelf-life of an edge?
- Are links between certain entity types many-many, many-one, one-one?

OSN Data collection techniques

- Ideally, the OSN would release full and accurate data to researchers directly.
- Privacy concerns and competitive reasons: OSNs will not do so
- Three general approaches used to collect data on OSNs:
  1. API driven
  2. Scraping-based
  3. Passive network measurement

API based approach

- Query the entities, properties and relationships.
- Must assume that the answers to queries via the API are up-to-date and accurate.
- API calls can produce random example of a particular type.
- Site might rate-limit number of API calls per user per day
- Order in which answers are supplied, limit on the answers may yield a highly biased sample

Scraping based

- Directly accesses the site via a Web client imitating the actions of a user
- Capture HTML, parse via a hand-crafted site-specific parser.
- More arduous than API-based methods and may still be throttled
- Scraper must contend with site redesigns which break the parser

Passive network measurement

- Sniff network traffic (at edge of a campus or enterprise network), sift out and parses requests to and from the OSN of interest
- Most honest view of the network in use as it captures properties of the network as its users experience it
- Significant privacy issues around sniffing and parsing individuals activities
- Many access modalities for modern OSNs (direct Web based, mobile Web, mobile app, external apps using API): hard to capture all accesses from any meaningful subpopulation of users.

Sampling alternatives

- Hundreds of millions of users; must fetch a few as sample and extrapolate and be aware of of biases
- Common sampling methods: based on a simple limited graph view of network
- Can't deal with dynamism of real-world OSNs (changes as it is sampled)
- Truly random sampling: need detailed knowledge of space of node ids, access arbitrary nodes given their id; or an API call to return a 'random' node (often far from random).
- Given OSN size, approach yields many isolated nodes: unsuitable for studying connectivity, reachability, or distance-based questions.
- BFS from few seed nodes: strongly influenced by the choice of initial nodes
- Tends to over-represent large connected components relative to islands and miss nodes with only outgoing links.

OSN properties of interest

Basic characteristics	Dynamic interaction	Network traffic specific	Social
Number of users	Inter-communication frequency	Protocol usage	Anonymity
Friend count distributions	Session duration	Induced overlay network	Privacy
Personal attributes	Diurnal properties	Byte-fraction distributions	
Communication options	Rate of change	Signature of individual OSNs	
Sub-communities	Popularity growth	Signature of intra-OSN functions	
Content diversity	External applications		
Ambient properties	Sub-session features		
Friendship link structure			

Facebook

- Papers of relevance to Facebook privacy: [Roo10, WXG11, MVGD10] (the more papers on this topic you read the more helpful it will be for your project)
- I will cover three of my papers in depth as it relates to privacy leakage and you will understand the techniques used to identify leakages.

[Roo10] Facebook Tracks and Traces Everyone: Like This!

[WXG11] Third-Party Apps on Facebook: Privacy and the Illusion of Control

[MVGD10] You are who you know: Inferring user profiles in Online Social Networks

Micro-content networks

- An average YouTube video is large, 10Mb or more
- Micro-content network messages are very small - typically consist of a couple of hundred bytes
- Short text line, location information, tinyURL
- Often a publisher-subscriber system with control on subscribers
- One to many communication possible, recipients can choose how to receive messages
- Twitter, Jaiku, Dodgeball (last two bought by Google and shut down), Utterz, LoopNote, Groovr
- Twitter has pretty much become the only game in town in micro-OSNs
- Over 200M users, billions of tweets daily

A few chirps about Twitter

- First paper to characterize Twitter, in 2008 (a million users)
- We used the API to do a constrained crawl (post-whitelisting)
- From crawl, learned about 67K users
- Twitter displays most recent tweets in public timeline on demand
- From timeline data learned about 36K users
- A tertiary crawl to validate methodology; not important here
- Now there is a Gardenhose interface to facilitate free data gathering while Firehose is for sale
- We learned about broadcasters, acquaintances, outlier users

References

[Kri09] B. Krishnamurthy. A measure of online social networks. 2009. Invited paper.

[KW08] Balachander Krishnamurthy and Craig E. Wills. Characterizing privacy in online social networks. In *Proceedings of the first workshop on Online social networks*, 2008.

[KW09] B. Krishnamurthy and C. Wills. On the leakage of personally identifiable information via online social networks. In *Proceedings of the Workshop on Online Social Networks*, August 2009.
`http://www.research.att.com/~bala/papers/wosn09.pdf`.

[KW10] Balachander Krishnamurthy and Craig E. Wills. Privacy leakage in mobile online social networks. In *Proceedings of the Workshop on Online Social Networks*, June 2010.
`http://www.research.att.com/~bala/papers/pmob.pdf`.

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

[MVGD10] Alan Mislove, Bimal Viswanath, Krishna P. Gummadi, and Peter Druschel. You are who you know: Inferring user profiles in online social networks. In *International Conference of Web Search and Data Mining (WSDM)*, February 2010. `http://www.ccs.neu.edu/home/amislove/publications/Inferring-WSDM.pdf`.

[Roo10] Arnold Roosendaal. Facebook tracks and traces everyone: Like this! Technical report, Tilburg Law School, November 2010. Legal Studies Research Paper Series No. 03/2011. `http://ssrn.com/abstract=1717563`.

[WCG+11] David Wetherall, David Choffnes, B. Greenstein, Seungyeop Han, Peter Hornyack, Jaeyeon Jung, Stuart Schechter, and Xiao Wang. Privacy revelations for web and mobile apps. In *Proceedings of*

HotOS, May 2011.
http://appanalysis.org/jjung/jaeyeon-pub/hotos2011-revelations.pdf.

[WXG11] Na Wang, Heng Xu, and Jens Grossklags. Third-party apps on Facebook: Privacy and the illusion of control. In *Proceedings of the ACM Symposium on Computer Human Interaction for Management of Information Technology*, December 2011. http://faculty.ist.psu.edu/xu/papers/conference/Wang_chimit_2011.pdf.