Privacy issues in the WiFi technology

Mathieu Cunche[†‡], Mohamed-Ali Kaafar[‡*], Roksana Boreli[*]

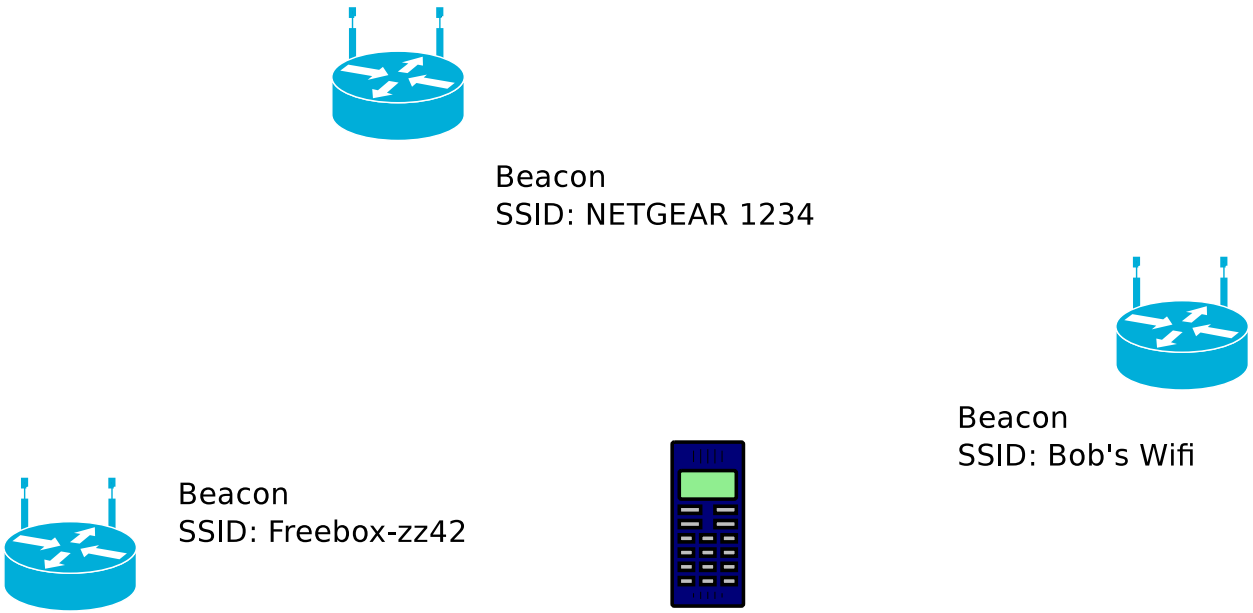
[†] INSA-Lyon/CITI Lab., [‡]INRIA, [*]National ICT Australia

Journées SEmba

Outline

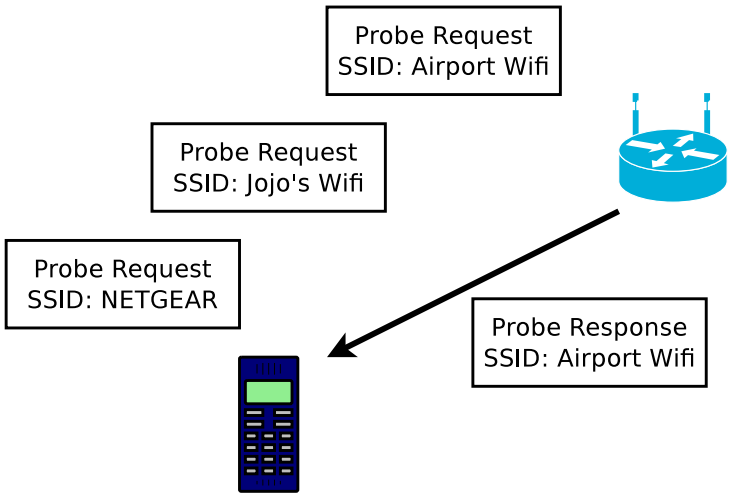
Wi-Fi service discovery I

- Passive service discovery mode
- AP broadcast Beacons
- Station listen to beacons and start connection when known SSID is detected



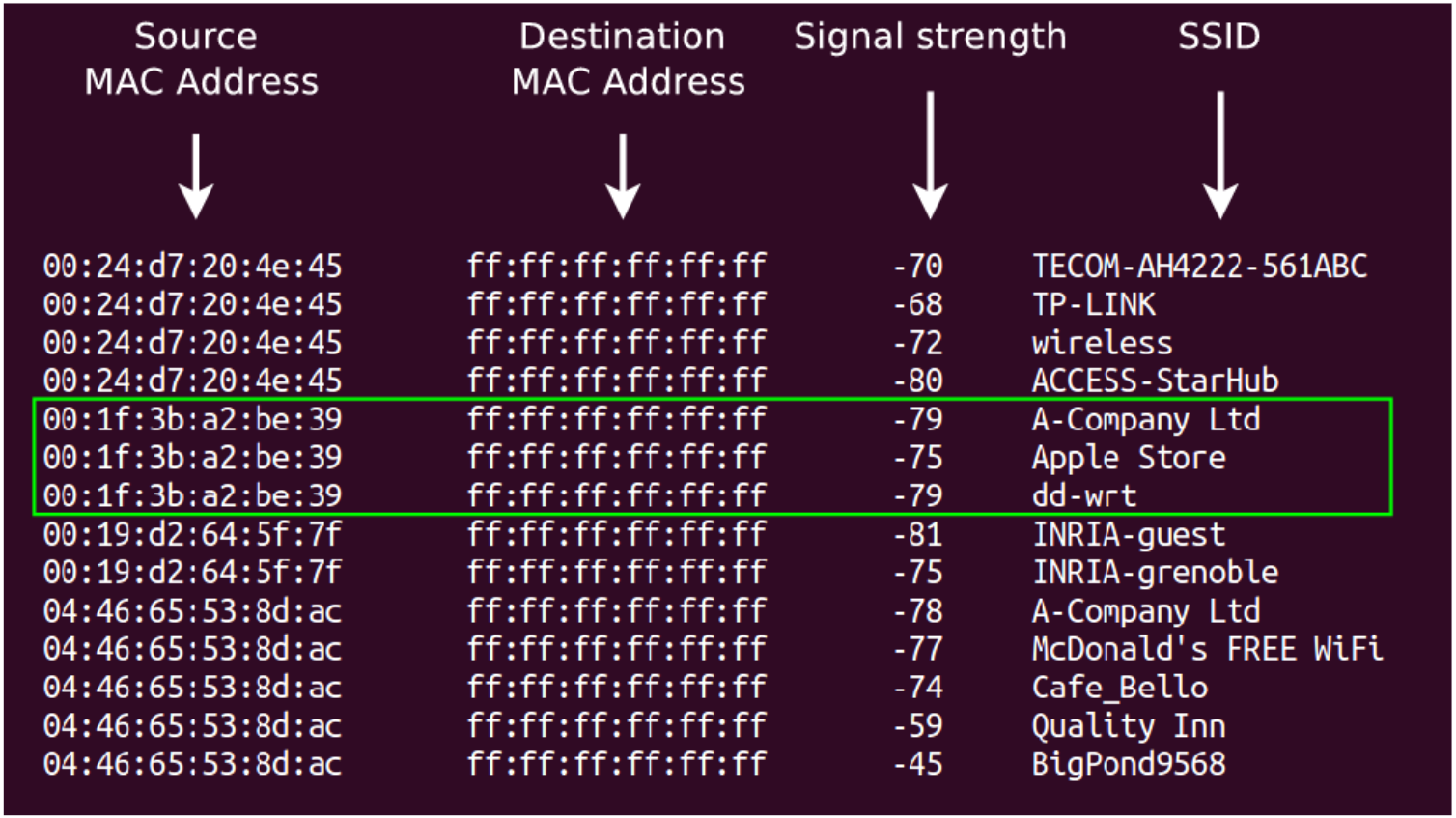
Wi-Fi Service Discovery I

- Wi-Fi Active service discovery mode
  - Stations probe for known Access Point (AP) in range
    - Probe request messages containing SSID of the AP
    - *Known* AP are stored in the Configured network list (CNL)



Wi-Fi Fingerprint

- Probe requests are broadcasted in plain text



Source MAC Address	Destination MAC Address	Signal strength	SSID
00:24:d7:20:4e:45	ff:ff:ff:ff:ff:ff	-70	TECOM-AH4222-561ABC
00:24:d7:20:4e:45	ff:ff:ff:ff:ff:ff	-68	TP-LINK
00:24:d7:20:4e:45	ff:ff:ff:ff:ff:ff	-72	wireless
00:24:d7:20:4e:45	ff:ff:ff:ff:ff:ff	-80	ACCESS-StarHub
00:1f:3b:a2:be:39	ff:ff:ff:ff:ff:ff	-79	A-Company Ltd
00:1f:3b:a2:be:39	ff:ff:ff:ff:ff:ff	-75	Apple Store
00:1f:3b:a2:be:39	ff:ff:ff:ff:ff:ff	-79	dd-wrt
00:19:d2:64:5f:7f	ff:ff:ff:ff:ff:ff	-81	INRIA-guest
00:19:d2:64:5f:7f	ff:ff:ff:ff:ff:ff	-75	INRIA-grenoble
04:46:65:53:8d:ac	ff:ff:ff:ff:ff:ff	-78	A-Company Ltd
04:46:65:53:8d:ac	ff:ff:ff:ff:ff:ff	-77	McDonald's FREE WiFi
04:46:65:53:8d:ac	ff:ff:ff:ff:ff:ff	-74	Cafe_Bello
04:46:65:53:8d:ac	ff:ff:ff:ff:ff:ff	-59	Quality Inn
04:46:65:53:8d:ac	ff:ff:ff:ff:ff:ff	-45	BigPond9568

- Wi-Fi Fingerprint = List of SSIDs broadcasted by a device

Privacy issues of service discovery I

- Active service discovery is bad for your privacy

- Allows tracking of individuals (MAC addr. broadcast)
- Wi-Fi fingerprint contains personal information

Privacy issues of service discovery II

- Personal information found in Wi Fi fingerprints
  - Link with a company/university
    INRIA-interne, *INSA-INVITE*, *GlobalCorp Ltd.*
  - Attended conferences
    SIGCOM-12, *Globecom11*
  - Visited places (hotel, restaurant, coffeeshop, airport)
    Hilton-NY WiFi, *Aloha Hotel WiFi*, *Brasserie de l'Est* ,*Sydney-airport-WiFi*
  - Individual's identity
    Marc Dupont's iPhone, *Bob Fhisher's Network*
  - Accurate geographic information
    Freebox-B4E781 $\rightarrow$ (-57.114,12.489)
  - Social links between individuals
    Onwers of [04:BB:48:11:74:F1] and [b8:FF:61:46:A5:E4] are friends

Link prediction

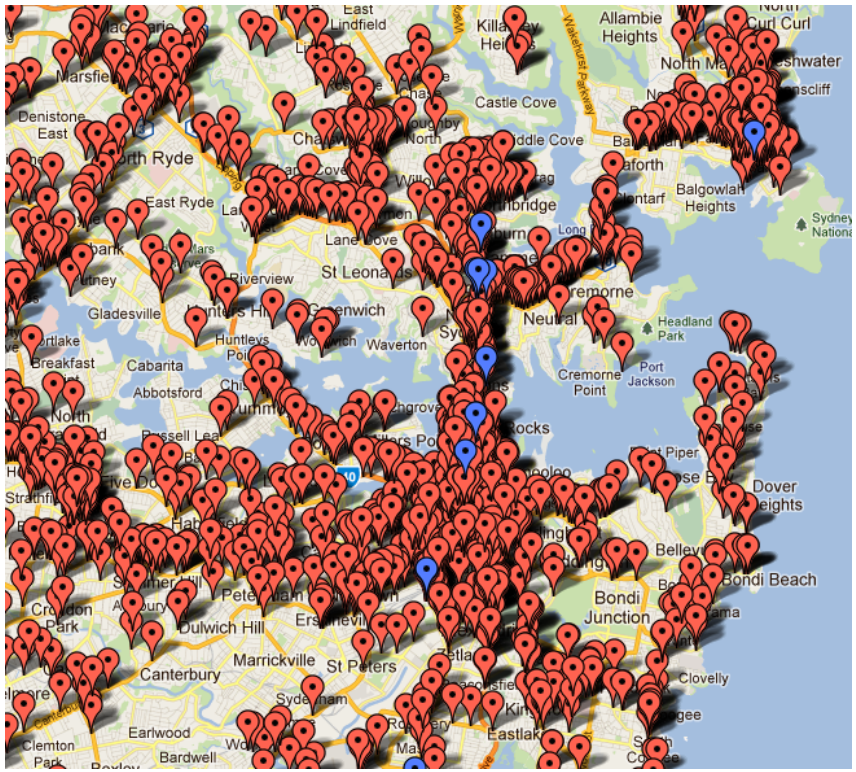
- The Link Prediction Problem: How to predict links between items ?
  - Within social/professional graphs, databases
- Prediction based on similarity between items
- Link prediction have been studied in several contexts
  - Based on shared friends [3]
  - Based on shared interests [1]
  - Based on temporal co-occurrences (contact length and frequency) [5, 2]
- Our idea : predicted links based on the Wi-Fi fingerprints
  - People with similar fingerprints are likely to be linked

Methodology

- Hypothesis : Wi-Fi fingerprint can reveal Links between individuals
  - Link prediction based on similarity between fingerprints
- Two data sets :
  - Controlled data set obtained from volunteers
    - Knowledge of the links
  - Wild data set collected in Sydney (8000+ devices, 24 000+ SSIDs)
    - Corpus to compute the frequency of SSIDs

Wild Data Set I

The Wild Data set



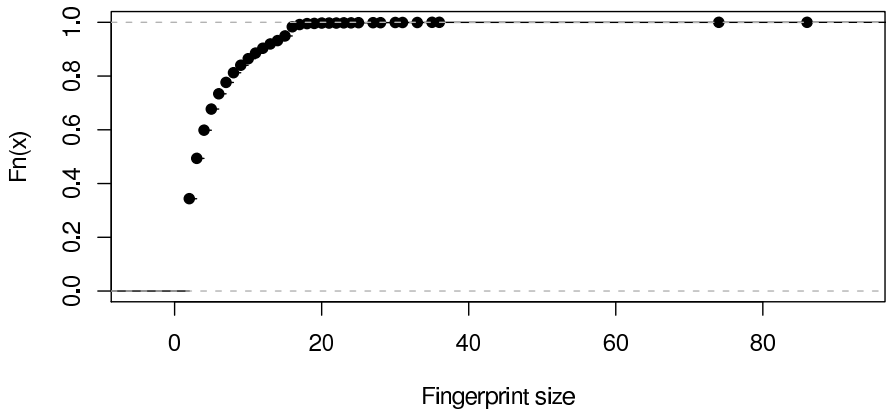
- WiFi fingerprints of 8000 devices, 24 000 SSIDs collected in Sydney over 5 months

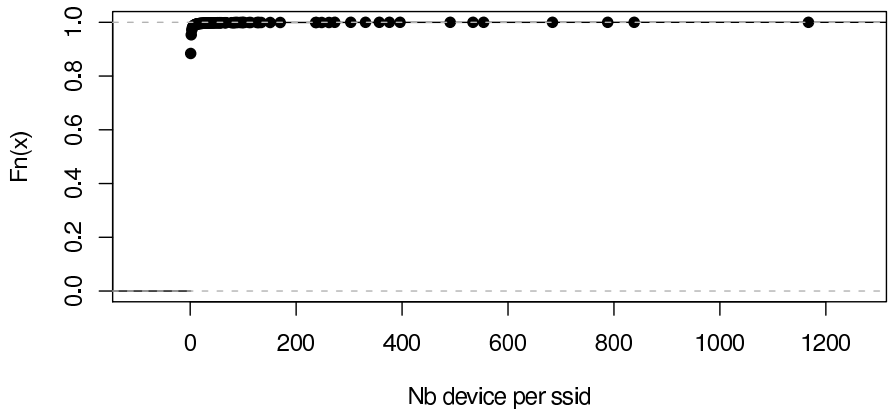
Wild Data Set II

Collecting 8000 WiFi fingerprints

- Hardware and software tools
  - A netbook + a WiFi interface
  - Monitor-mode enable drivers
  - Network traffic tools (wireshark)
- Harvesting the data
  - Walk the streets with the netbook in your backpack
  - Collect probe requests broadcasted by surrounding phones
  - Estimated range: 20-30 meters

Wild Data Set III

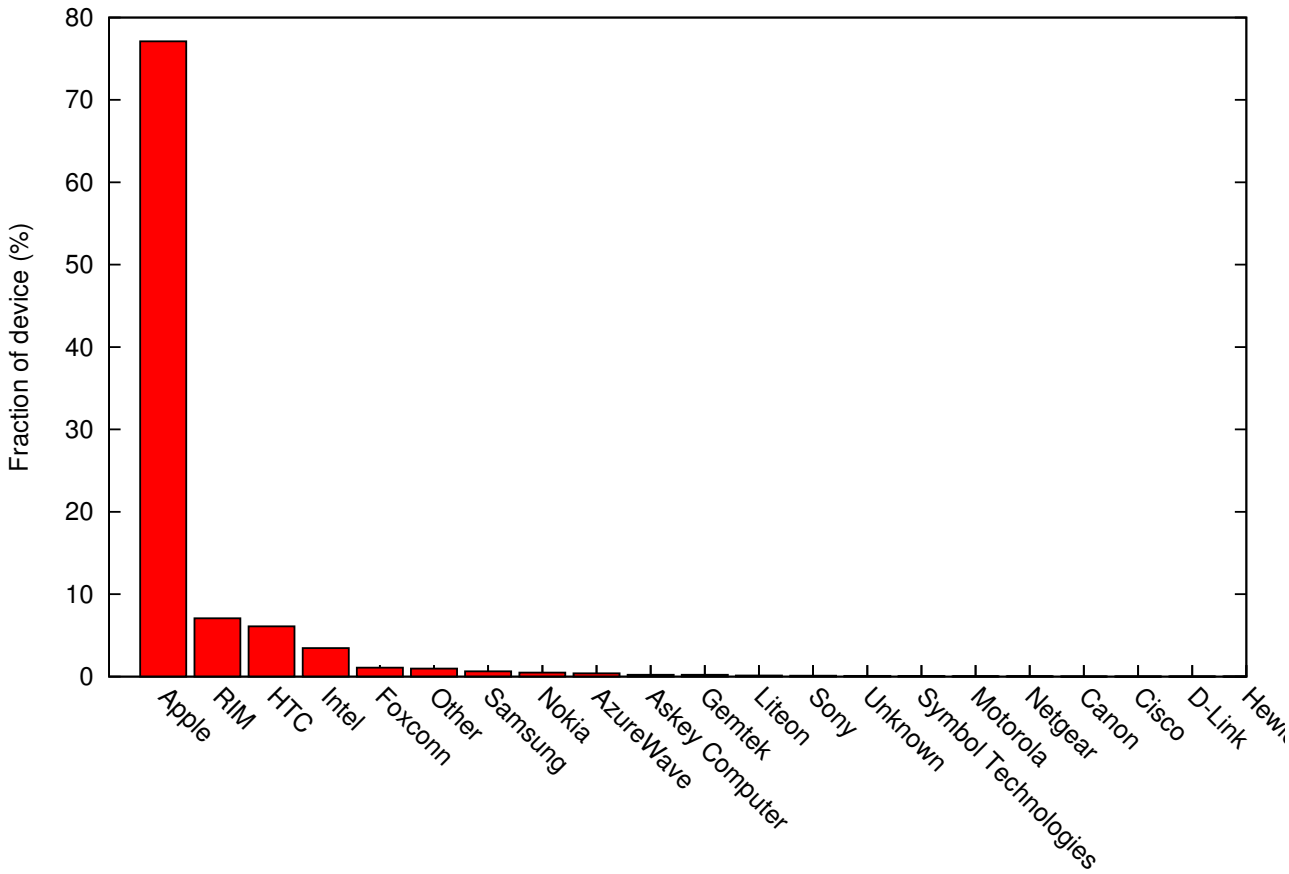




- Fingerprints size : between 1 and 80 SSIDs
- Some SSIDs are common
  - *NETGEAR* (838 devices), *McDonald's FREE WiFi* (491 devices)
- Other SSIDs are rare
  - *BigPondC8EEE5* (1 device), *John Doe's Network* (1 device), *2012 is the end of world?* (1 device), *mercure-ibis-brisbane* (2 devices)

Interface vendor I

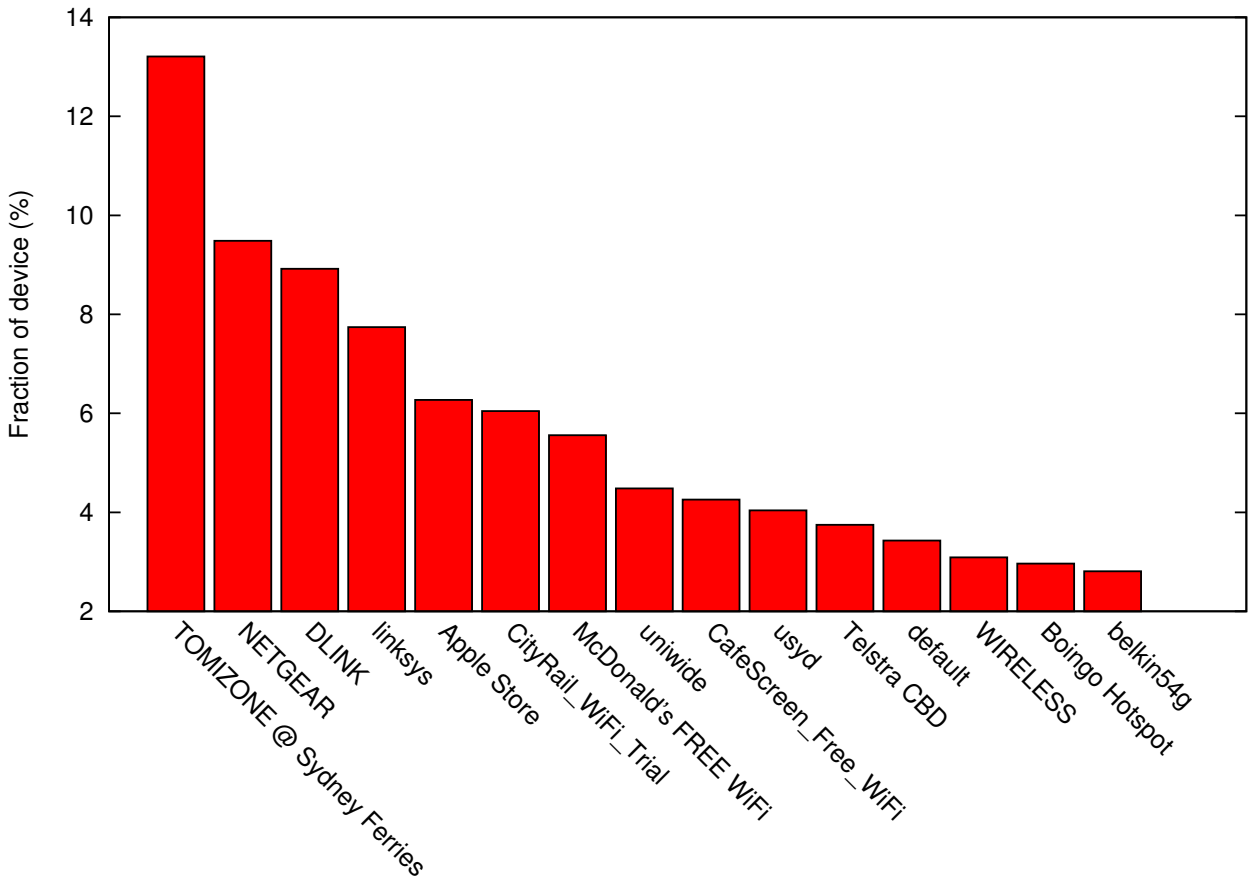
- MAC Address reveal the ID of the interface *manufacturer*



- Apple devices are very chatty

Popular SSID I

- Top most frequent collected SSIDs



- Default router names and shop/restaurant hotspots

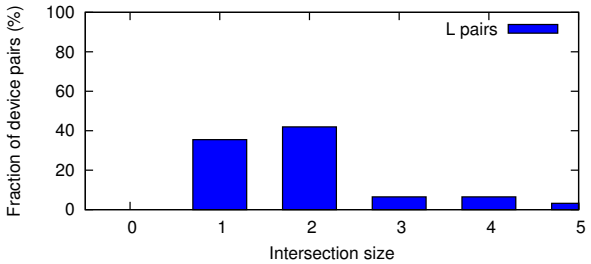
Outline

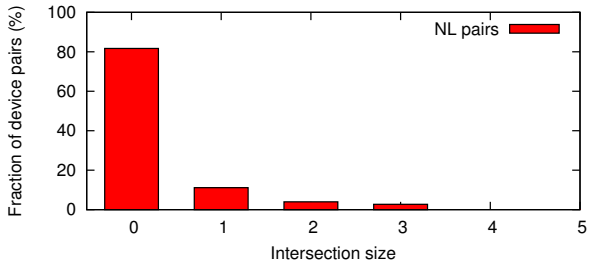
Device linkability

- Linking devices
  - Similarity between fingerprints reflect a link between users
  - People living/working together tends to have AP in common
- A controlled data set
  - Fingerprint collected from a group of volunteers
    - 30 existing strong social links
  - Existence of link is known for each pair of volunteers
    - Two class of pairs: Linked pairs and Non-Linked pairs

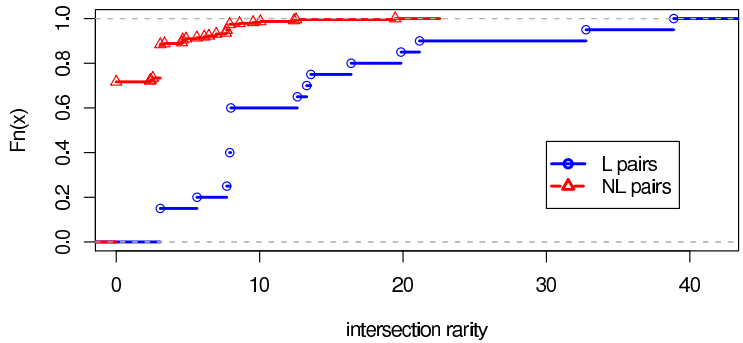
Fingerprint pairs characteristics I

- Fingerprint intersection size and rarity of Linked and Non-Linked pairs





$$\text{Rarity}(X, Y) = \sum_{z \in X \cap Y} -\log f_z$$



- Linked pairs have intersection with more and less frequent elements than Non-Linked pairs

Fingerprint similarity I

Conclusions on the design of the similarity metric

- Both the number and frequency of shared SSIDs should be considered
  - Number of shared SSIDs
    - How many network in common
  - Frequency of shared SSIDs
    - How common are these networks names *McDonalds Free WiFi* vs. *Max Power's WiFi*

Fingerprint similarity I

Conclusions on the design of the similarity metric

- Both the number and frequency of shared SSIDs should be considered
  - Number of shared SSIDs
    - How many network in common
  - Frequency of shared SSIDs
    - How common are these networks names *McDonalds Free WiFi* vs. *Max Power's WiFi*

Similarity metrics I

- Considered Similarity metrics
  - Cosine-IDF and Jaccard index

$$\text{Cosine-idf}(X,Y) = \frac{\sum_{x \in X \cap Y} \text{idf}_x{}^2}{\sqrt{\sum_{x \in X} \text{idf}_x{}^2}\sqrt{\sum_{y \in Y} \text{idf}_y{}^2}} \qquad \text{J}(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

where idf_x : inverse document frequency of x

  - Adamic [1], modified Adamic

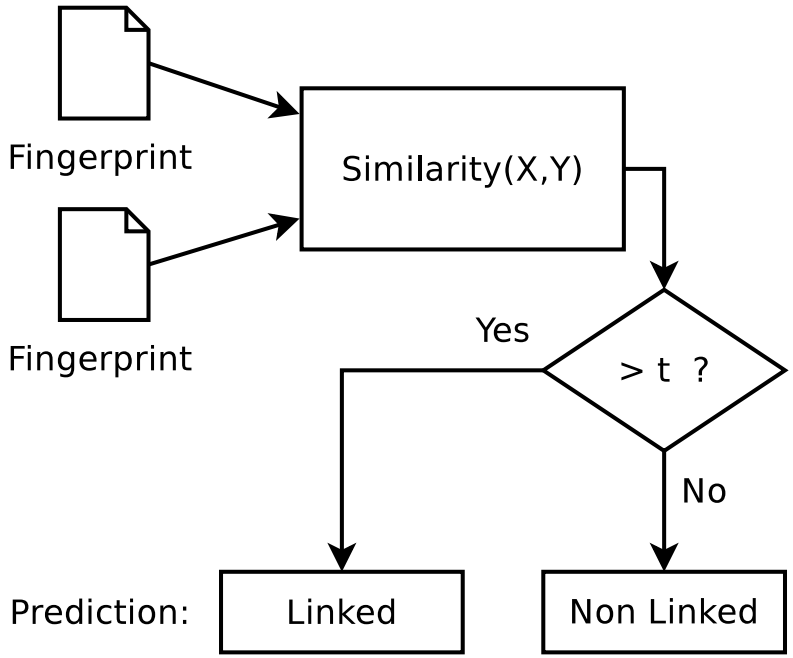
$$\text{Adamic}(X,Y) = \sum_{x \in X \cap Y} \frac{1}{\log f_x} \qquad \text{Psim-}q(X,Y) = \sum_{x \in X \cap Y} \frac{1}{f_x^q}$$

where f_x : document frequency of x

- The higher the similarity the more likely a link exists

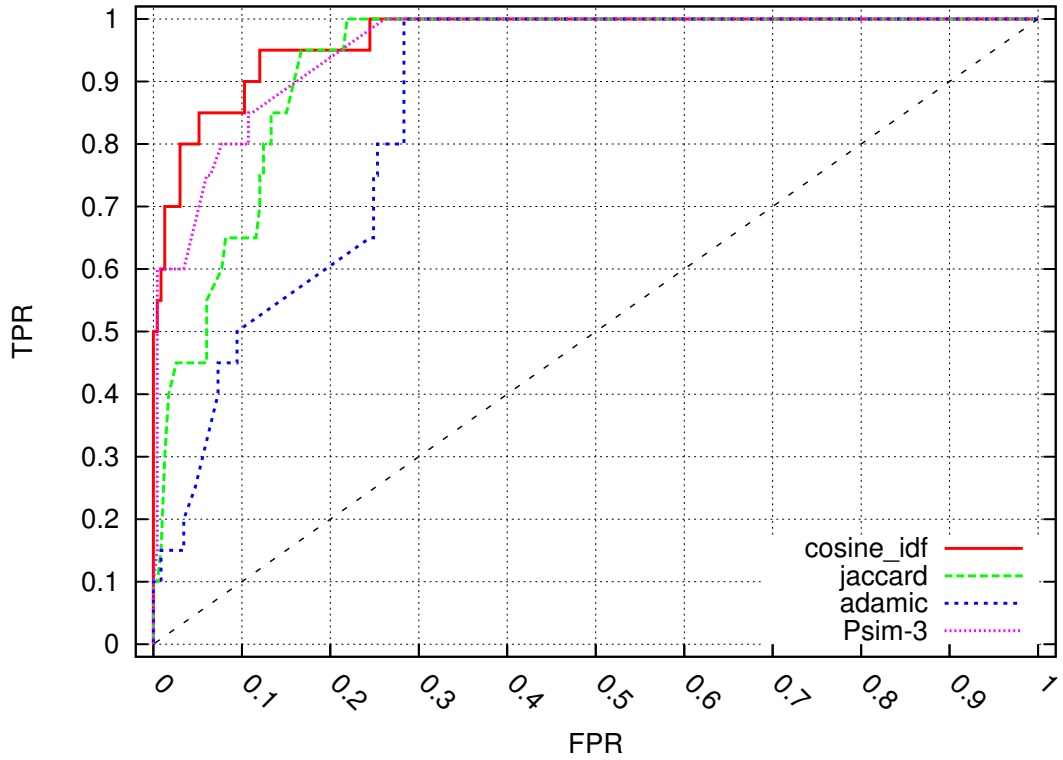
Similarity metrics II

- Classifier based on similarity metric
  - Similarity score compared to a threshold



Evaluation

- Controlled data set used to test performances
  - True positive rate (TPR) vs. False positive rate (FPR)



- Best metrics: Cosine-IDF and modified Adamic (Psim-3)

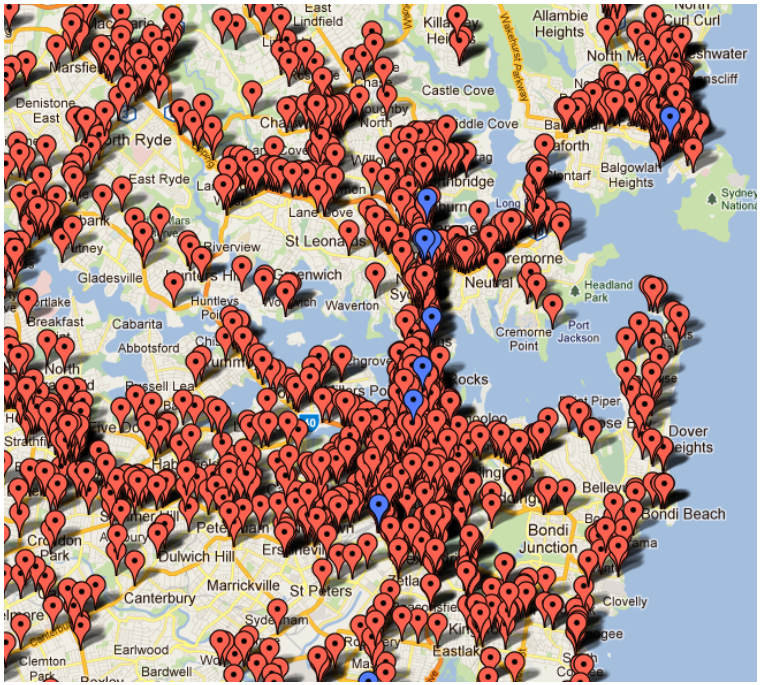
Geolocation information I

From SSIDs to geolocation information

- Wireless network databases
  - WiFi-based Geolocation
    - Submit BSSID of surrounding WiFi APs, get geolocation coordinates
    - Alternative to GPS
    - Service provided by Google, Apple, Skyhook
  - Databases maintained by hobbyist
    - Crowdsourced data (smartphone app.)
    - Extensive information about AP: BSSID, SSID, encryption, geoloc, open/closed ?
    - Examples: Openbmap, WiGle

Geolocation information II

- Combining the data
  - Join the WiFi fingerprint with the geoloc. databases
  - Each device is now associated to a set of geolocation coordinates
  - Reveal where you live/work/travel/...



Geolocation information III

- Limitations
  - Only hobbyist databases support SSID lookup
  - The largest databases (Google) only support BSSID lookup
  - Some SSID match large number of scattered geolocation (McDonalds WiFi)
  - Some SSID are missing from those databases

Countermeasures

Possible countermeasures

- What you can do
  - Disable active service discovery
  - Delete outdated configured networks
  - Turn off WiFi whenever possible
- What the manufacturer can do
  - Implement privacy preserving active service discovery [4]
  - Use blind probe request
  - Provide clear configuration options

Countermeasures

Geolocation and WiFi service discovery

- Remark on WiFi networks
  - Access Points cover a limited area (house, building, campus, ...)
  - No need to probe for a network if know we are kilometers away from it
- A geolocation assisted active discovery mode
  - Record the location of configured AP
  - Only probe for network located next to my current position
- Effects on privacy
  - Reveal only a part of the WiFi fingerprint
  - Broadcasted SSIDs gives little information (close to the corresponding AP)

Outline

Conclusion

- Your Wi-Fi device leaks private information
  - Information broadcasted in plain text
  - Social links, visited places, identity ...
- Potential applications
  - Forensic: identify the members of a criminal network
  - Marketing and targeted advertisement
  - Physical Analytics
- 802.11 standards privacy tooks years to be considered
  - First 802.11 standard introduced in 1999
  - Wi-Fi privacy issues have been noticed few years ago (2007)
- Too late to be fixed ?
  - Millions of devices and AP already deployed

Bibliography I

Lada A. Adamic and Eytan Adar.
Friends and Neighbors on the Web.
SOCIAL NETWORKS, 25:211–230, 2001.

David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon M Kleinberg.
Inferring social ties from geographic coincidences.
Proceedings of the National Academy of Sciences of the United States of America, 107(52):22436–22441, 2010.

David Liben-Nowell and Jon Kleinberg.
The link prediction problem for social networks.
In *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, pages 556–559, New York, NY, USA, 2003. ACM.

Janne Lindqvist, Tuomas Aura, George Danezis, Teemu Koponen, Annu Myllyniemi, Jussi Mäki, and Michael Roe.
Privacy-preserving 802.11 access-point discovery.
In *Proceedings of the second ACM conference on Wireless network security*, WiSec '09, pages 123–130, New York, NY, USA, 2009. ACM.

Daniele Quercia and Licia Capra.
Friendsensing: recommending friends using mobile phones.
In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 273–276, New York, NY, USA, 2009. ACM.