**Privacy-preserving Wi-Fi tracking systems**

Célestin Matte¹, Marine Minier¹, Mathieu Cunche¹, Franck Rousseau²

¹Université de Lyon, INRIA, INSA-Lyon, CITI-INRIA, France
²Grenoble Institute of Technology, CNRS Grenoble Informatics Laboratory UMR 5217, Grenoble, France

---

**Wi-Fi TRACKING SYSTEMS**

- Wi-Fi-enabled smartphones frequently emit frames because they scan for access points
- Each frame contains a unique identifier of the device: the MAC address
- Wi-Fi tracking systems monitor and store these frames in order to make statistics:
  - number of visitors
  - frequency of visits
  - travel paths
  - etc.

---

**PRIVACY ISSUES**

- Lot of private information stored - much more than needed to make basic statistics
- Anyone with access - legitimate or not - to that data can get a lot of information about passers-by
- In France, made illegal by the “Informatique et Liberté” law (LIL). The CNIL made precise recommendations for this case of information collecting so that one knows if they respect the law [1]:
  - Data must be deleted when the person exits the place
  - Used algorithm must ensure a strong collision rate, i.e. an identifier must correspond to several people
  - People must give their explicit consent if one wants to store data for a longer period (opt-in system).
- We can work on storing only information valuable for statistics
- There is no miracle solution
- We aim to propose a privacy/utility tradeoff

---

**OBJECTIVES**

- Short term:
  - First, comparing the different methods
  - Finding a good privacy metric
- Long term:
  - Having a ready-to-use system with the same functionalities than existing Wi-Fi tracking systems, augmented with privacy-by-design

---

**DATA COLLECTION**

- We need datasets to perform all our tests, containing logs of raw MAC addresses seen at a certain time at a certain place.
- Legislation forbids the collection of such datasets, as they contain personal identifiers
- Need consent of all concerned people
- Difficult to set up
- Possible solutions:
  - getting the consent of concerned people (one try at ACM Middleware, small dataset obtained)
  - generate synthetic datasets from models

---

**PRIVACY METRICS**

- We need metrics to quantify “privacy”, in order to evaluate our methods
- Possible metrics are (non-exhaustive list):
  - K-anonymity
  - Collision rate: how many devices may be identified by the same identifier?
  - Entropy?
  - Other considerations:
    - Is the collision rate evenly distributed?
    - What happens on extreme values? (e.g. very few number of passers-by)
- We have to determine which one (among those ones or other ones) best fits ours needs

---

**HASHING**

- Principle: Using a simple cryptographic hash function, such as MD5 or functions of the SHA family, with no salt
- Problems:
  - Easy to reverse in the case of MAC addresses [2]
  - Almost no collision
  - Almost useless

---

**HASHING AND TRUNCATION**

- Principle: Same as hashing, but only keep a small part of the result
- the less you keep, the more collisions you’ll get
- Also possible to truncate before hashing, in order to manually increase collision rate (first 3 bytes of the MAC address are the constructor identifier)
- Problem: Still easy to reverse?

---

**BITMAPS**

  
  $B(x) = \sum_{i=0}^{k-1} \lfloor \frac{x}{2^i} \rfloor$

- Problems:
  - Unequal collision probability (half MAC addresses will set first bit on while only one will set the last but to one)
  - Very approximate for low numbers of MAC addresses (which is our case most of the time)

---

**BLOOM FILTERS**

- Data structure useful for many applications
- Insertion and search in $O(1)$
- ... but false positives
- ... which we use for privacy purpose
- We aim to be able to say that a person may have been here, but never be sure about it
- Principle: about the same as bitmaps, but use several hashing functions spreading bits uniformly on the array of bits

---

**OTHER POSSIBILITIES**

- Using other phone characteristics instead of MAC addresses (plenty of fingerprinting techniques exist) [5]
- Hashing with salts (keys), and destroying keys after a predefined period of time. Key destruction is not a trivial problem.

---

**FUNDING**

- Project founded by Academic Research Community (ARC), Rhône-Alpes region (ARC 7).
- Partnership with UrbanLyon
- Searching for partners with an infrastructure to share (electricity + network access)

---

**REFERENCES**

[1] CNIL’s obligations
[2] Demir, Levent and Cunche, Mathieu and Lauradoux, Cédric
Analysing the privacy policies of Wi-Fi trackers and statistics.
Mathematical correction for fingerprint similarity measures to improve chemical retrieval
Probabilistic counting
[5] Xu, Qiang and Zheng, Rong and Saad, Walid and Han, Zhu
Device Fingerprinting in Wireless Networks: Challenges and Opportunities

---

---

---