Privacy-preserving Wi-Fi tracking systems

Célestin Matte, Marine Minier, Mathieu Cunche, Franck Rousseau

▶ **To cite this version:**

Célestin Matte, Marine Minier, Mathieu Cunche, Franck Rousseau. Privacy-preserving Wi-Fi tracking systems. Citi Lab PhD day, Apr 2016, Villeurbanne, France. , 2015. hal-01535820

HAL Id: hal-01535820
https://hal.science/hal-01535820v1

Submitted on 13 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

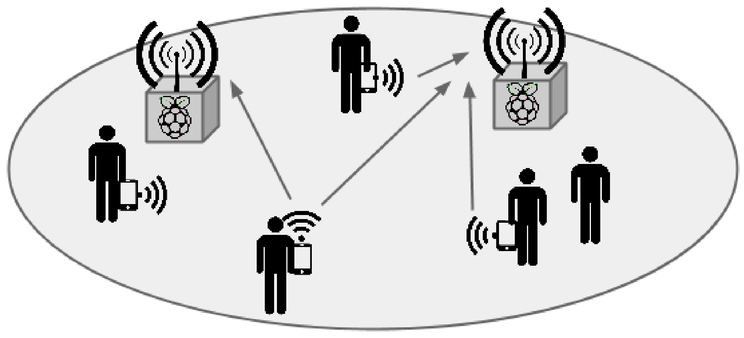
Privacy-preserving Wi-Fi tracking systems

Célestin Matte[1], Marine Minier[1], Mathieu Cunche[1], Franck Rousseau[2]

[1]Université de Lyon, INRIA, INSA-Lyon, CITI-INRIA, France
[2]Grenoble 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 recommandations 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 it 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

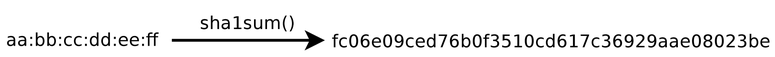
- We need datasets to perform all our tests, containing logs of raw MAC addresses seen at a certain time at a certain place.
- Legislation forbids the collection of such datasets, as they contain personal identifiers
- Need consent of all concerned people
  - difficult to set up
- Possible solutions:
  - 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 these ones or other ones) best fits ours needs

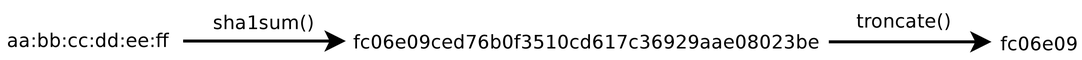
Hashing

- 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



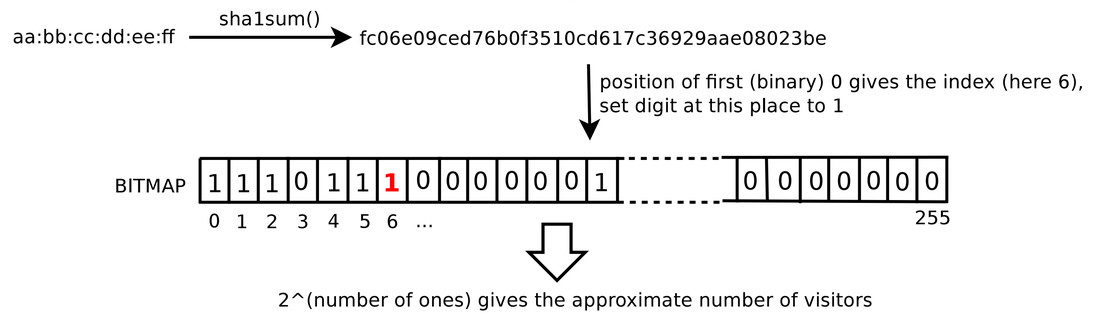
Hashing and troncation

- 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 troncate 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

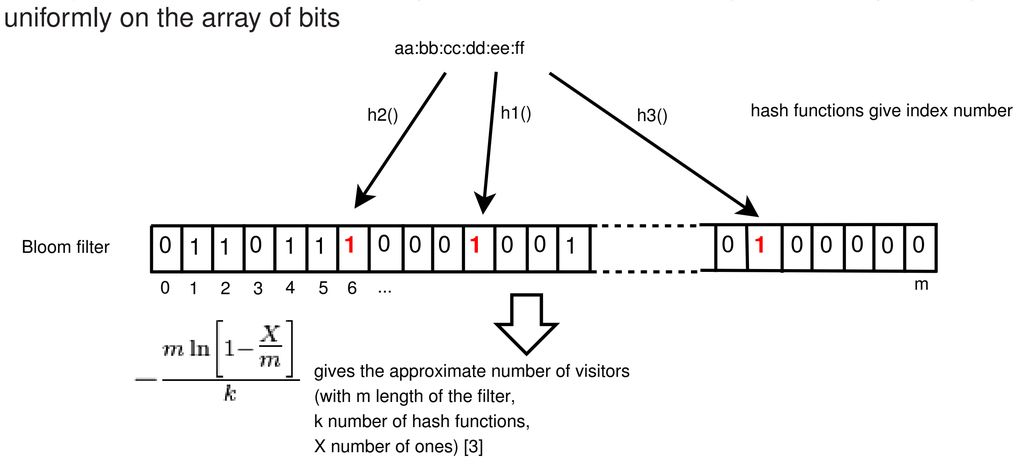
- Principle: logarithmic repartition of the hashes [4]



- Problems:
  - Unequal collision probability (half MAC addresses will set first bit ot one while only one will set the last bast 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



- Problems:
  - It may be possible to know for sure that some MAC addresses were seen
  - May still give bad results for low numbers of MAC addresses (tests ongoing)

Other possibilities

- Using other phone characterisitics 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.

Deployment

- Simple tools: raspberry pis with Wi-Fi dongles

- Partnership with UrbaLyon
- Searching for partners with an infrastructure to share (electricity + network access)

Funding

- Project founded by Academic Research Community (ARC), Rhône-Alpes region (ARC 7).

References

[1] CNIL’s obligations
http://www.cnil.fr/linstitution/actualite/article/article/mesure-de-frequentation-et-analyse-du-comportement-des-consommateurs-dans-les-magasins/, accessed on 2015.04.07.

[2] Demir, Levent and Cunche, Mathieu and Lauradoux, Cédric
Analysing the privacy policies of Wi-Fi trackers and statistics.

[3] Swamidass, S. Joshua; Baldi, Pierre (2007)
Mathematical correction for fingerprint similarity measures to improve chemical retrieval

[4] Flajolet, Philippe and Martin, G Nigel
Probabilistic counting

[5] Xu, Qiang and Zheng, Rong and Saad, Walid and Han, Zhu
Device Fingerprinting in Wireless Networks: Challenges and Opportunities