Privacy-preserving Wi-Fi Analytics

Barcelona, Spain
PETS 2018

Mohammad Alaggan[⋆] Mathieu Cunche[†] Sébastien Gambs[‡]

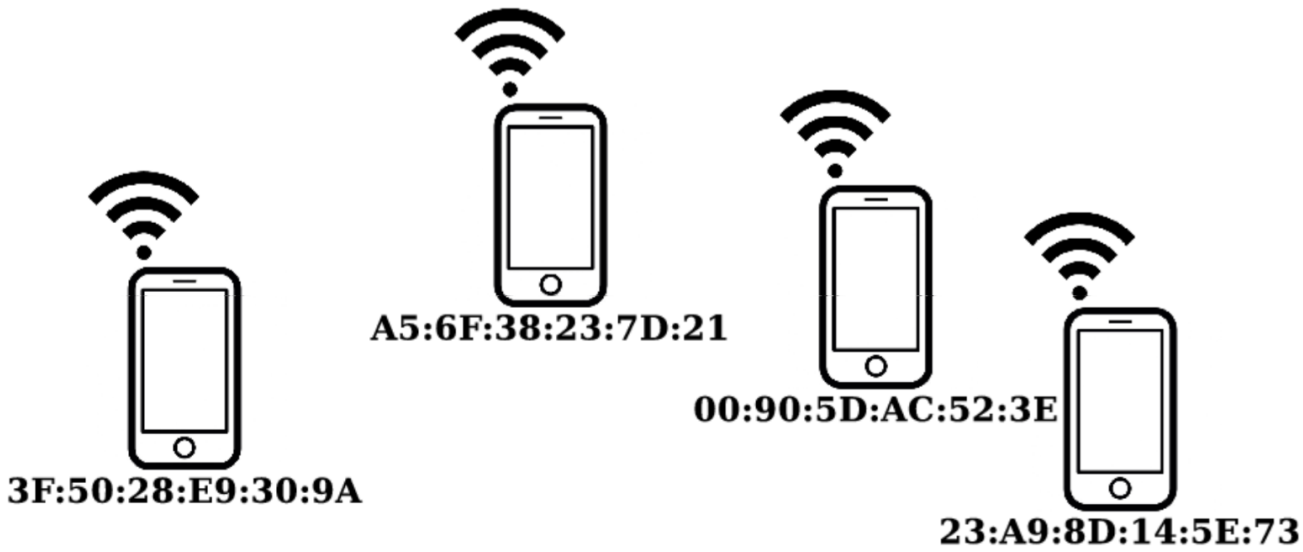
[⋆] Antidot, France (Work done while at Inria Lyon, France)
[†] Univ Lyon, Inria, France
[‡] Université du Québec à Montréal, Canada
mohammad.nabil.h@gmail.com

July 25, 2018

Context

Wi-Fi devices as personal beacons

- Wi-Fi enabled devices broadcast a **unique** ID: the **MAC address**
  - Connected: in Data, Management and Control Frames
  - Disconnected: in probe-requests (Management) Frames



Physical Analytics

- **Objective**: Measure and analyse human activity through Wi-Fi
  - One MAC address = One person
- **Examples of analystics tasks**:
  - Number of visitors
  - Duration/frequency of visits
  - Most popular paths between different locations
  - ...

source : Libelium

Current industrial practices for protecting privacy are not good enough

- Most of the companies rely on hashing to prevent the re-identification of the MAC address
- Hashes can be reversed in minutes using brute-force attack [DCL'14]

Time	Location	MAC
12:09	A-4	00:11:11:11:11:11
12:12	B-4	00:11:11:11:11:11
12:13	E-5	00:22:22:22:22:22
12:13	F-4	00:33:33:33:33:33
12:14	B-4	00:11:11:11:11:11

$\rightarrow$

Time	Location	Hash (md5)
12:09	A-4	fb2d5084c0ad1fdf6c29fe2aa323b758
12:12	B-4	fb2d5084c0ad1fdf6c29fe2aa323b758
12:13	E-5	69dc015b56448651561e1a4301ac9b4d
12:13	F-4	07024831442e8b86a06e905fd4d391ce
12:14	B-4	fb2d5084c0ad1fdf6c29fe2aa323b758

[DCL'14] L. Demir, M. Cunche, and C. Lauradoux. **Analysing the privacy policies of Wi-Fi trackers**, WPA'14

Threat model (Pan-Privacy [DNPRY'10])

- Attacker: **internal actor** (data collector) or **external intruder**
- Resource to protect: **internal state** of the system *and* the final output
- Protection must be done on-the-fly, as each MAC address is observed

C. Dwork, M. Naor, T. Pitassi, G. N. Rothblum, and S. Yekhanin. **Pan-Private Streaming Algorithms**. ICS'10

Pan-Privacy

Pan Privacy (informal and simplified) [DNPRY'10]

An algorithm is ε-*differentially pan-private* if the distribution of **both**:

- The internal state of the algorithm
- The final output

does not differ too much (depending on ε) if one MAC address was added

- Intention: from the internal state of the system and the output, the adversary cannot distinguish whether or not the MAC address of the user is present in the encoded set

[DNPRY'10] C. Dwork, M. Naor, T. Pitassi, G. N. Rothblum, and S. Yekhanin. Pan-Private Streaming Algorithms. ICS'10

Approach

Observation

Many mobility analytics can be based upon a primitive:
Cardinality Set Operations
(Also known as **Count-Distinct Queries**)
between different locations at different times

Example (Mobility Analytics)

	Temporal	Spatial	Set Operation
Number of visitors			Cardinality
Number of visitors		✓	Union
Amout of time they spend	✓		Intersection
Frequency of their visits	✓		Intersection
Their movement trajectories	✓	✓	Intersection
Most frequently taken path	✓	✓	Intersection

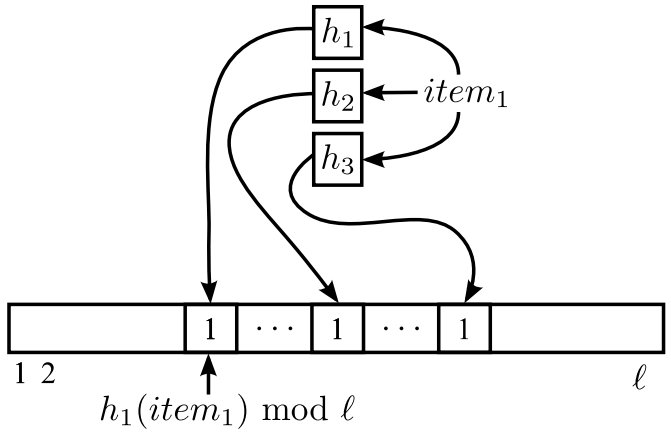
Our Approach

- **Key idea**: design a privacy-preserving data structure for computing the **Cardinality Set Operations** while protecting the **privacy** of individual users
- Agnostic to data source (**not limited to Wi-Fi**)
  - Cellular-based mobility analytics (Call-Detail-Records) [1]
  - Web analytics
  - Any system with unique identifiers...
- **Designed data structure**: based on **Bloom filters** that are perturbed to ensure **differential privacy** and built on the fly to ensure **pan-privacy**.
- **Non-interactive**: create the data structures first, specify the mobility analytics to compute later
- **Decentralized**: No need to coordinate between sensors

[1] [AGMT'15] Alaggan M., Gambs S., Matwin S., Tuhin M., **Sanitization of Call Detail Records via Differentially-Private Bloom Filters**. DBSec 2015

Bloom Filters [Bloom 1970]

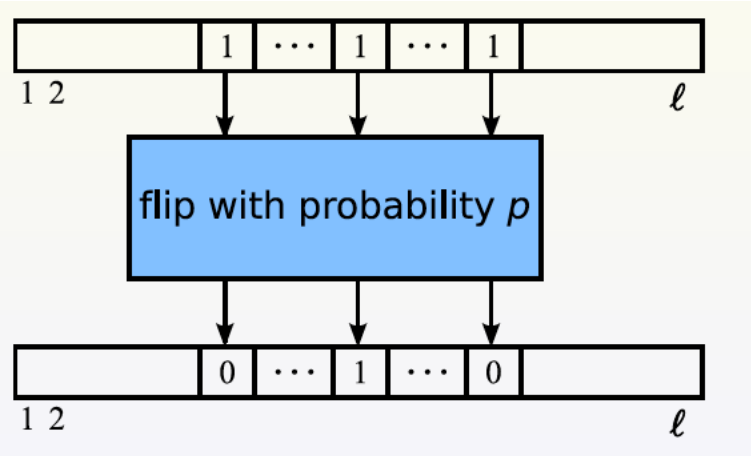
- Sets can be represented as Bloom filters



- Two operations: **insert** and **contains**
- Highly efficient in space and time
- Small probability of false positives, no false negatives
- Can add but cannot remove elements
- Not private: can be exhaustively queried

BLIP [AGK 12]

- Bloom Filter with **Differential Privacy** guarantees
- BLIP = BLoom-then-flIP
  - **Step 1**: Represent a set of identifiers as a Bloom filter
  - **Step 2**: flip each bit indepdendently and identically at random with probability $p < 0.5$.



- Estimator for distinct number of stored identifiers [BFG'14]

[BFG'14] Balu R., Furon T., Gambs S., **Challenging differential privacy: the case of non-interactive mechanisms**. In ESORICS 2014

Pan-Private BLIPs

- Choose two Bernoulli distributions, $D_0 \neq D_1$, according to ε

Pan-Private BLIP: Initialize

- Initialize all bits randomly from D_0

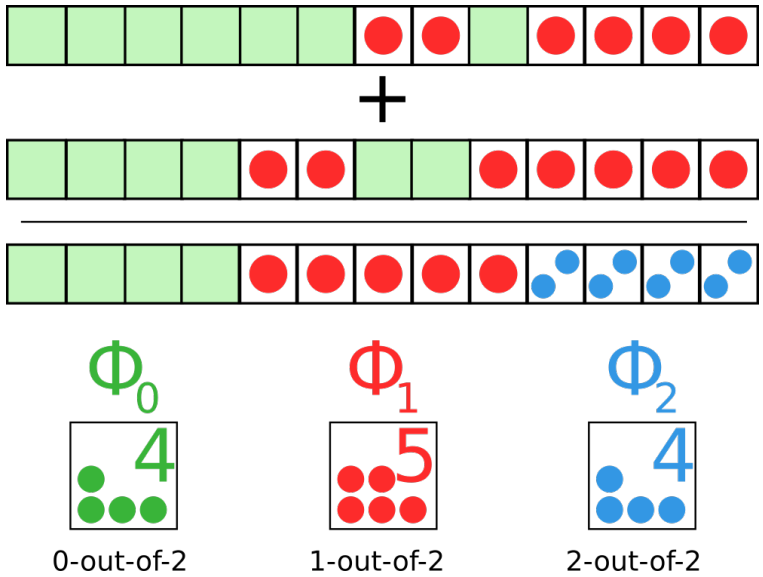
Pan-Private BLIP: Add element x

- Set bits $h_1(x), h_2(x), \ldots, h_k(x)$ randomly from D_1

Distinct-Count Queries for n BLIPs

Example (1/2) : Plain (unflipped) Bloom filters

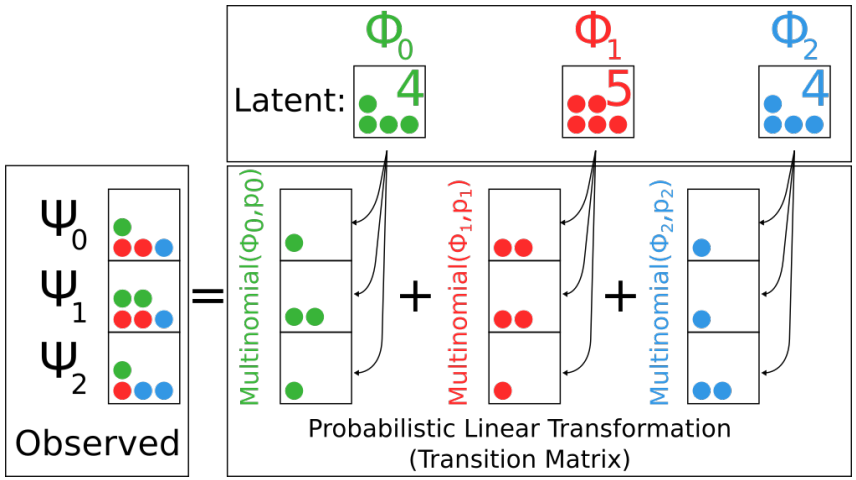
- Given two **unflipped** Bloom filters of size m
- Add them component-wise (over the integers)
- Tally the components
- Intersection ≈ 4 (number of components of count 2)
- Union ≈ 9 (number of components of count ≥ 1)



Distinct-Count Queries for n BLIPs

Example (2/2) : Pan-Private BLILPs

- Given two **flipped** Bloom filters of size m
- Add them component-wise (over the integers)
- Tally the components
- Estimate the unflipped tally [ACM 17]

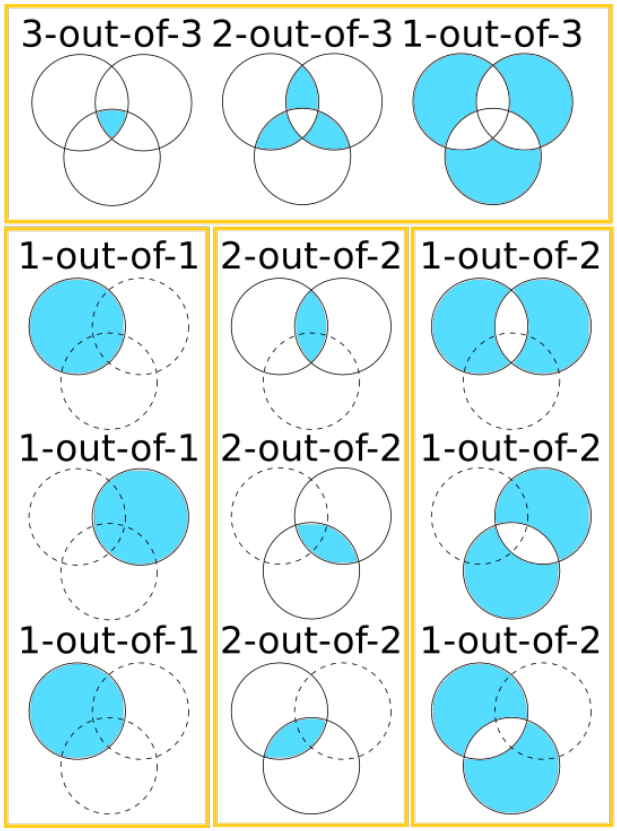


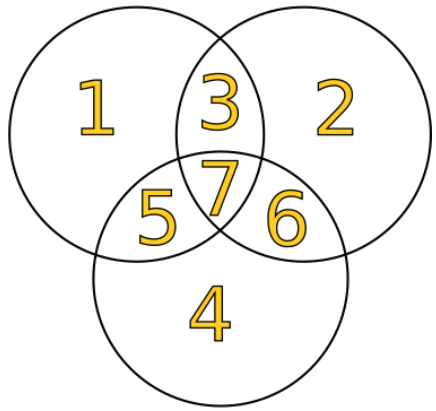
Distinct-Count Queries for *n* BLIPs

The general case: Symmetric Counts (t-out-n counts)

Number of elements belonging to **exactly** t sets out of n

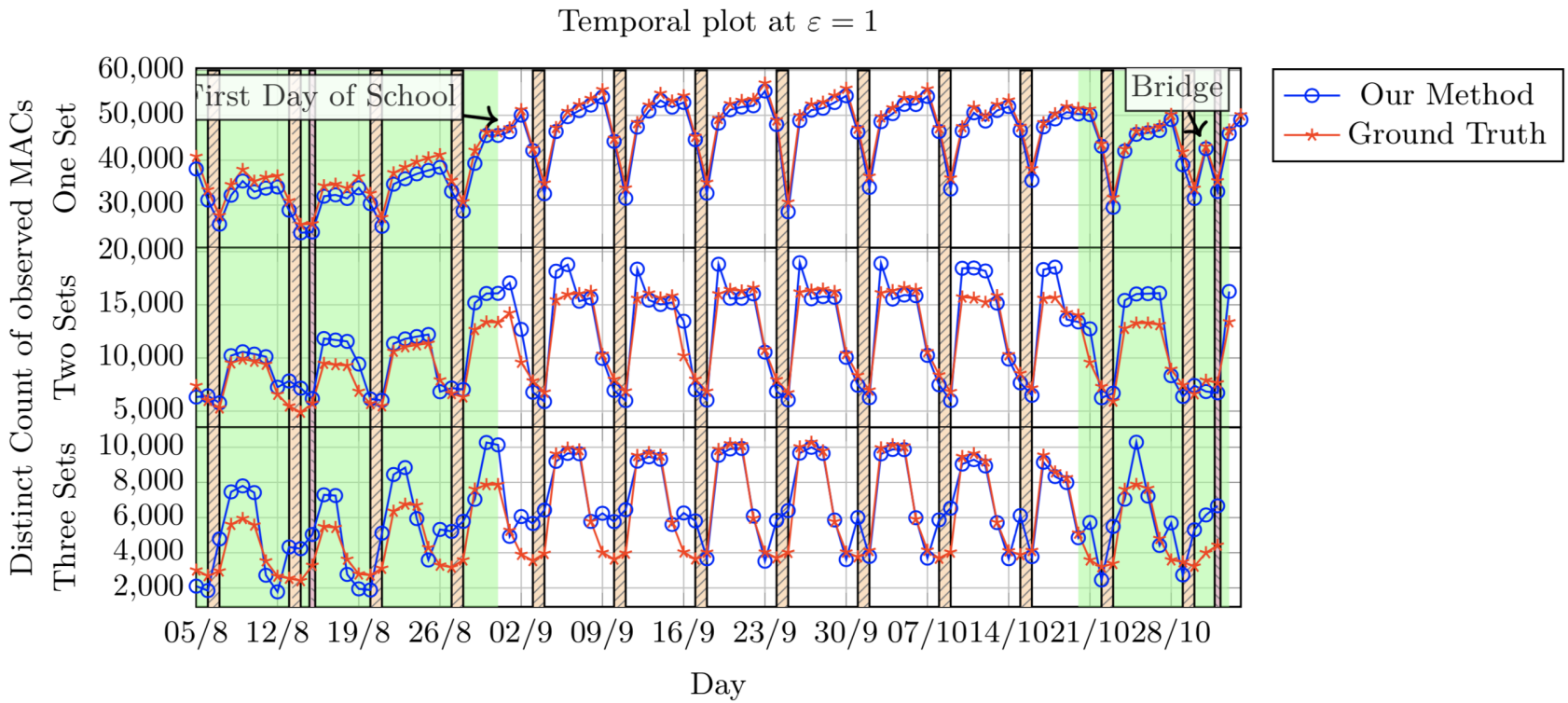
- Can estimate any count from several symmetric counts





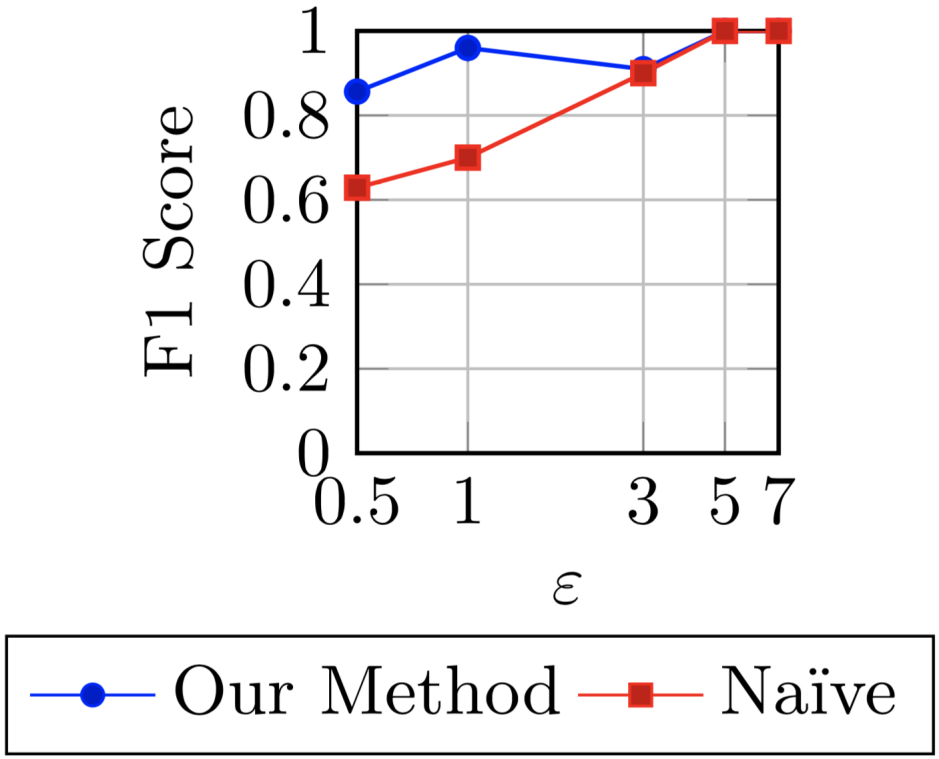
Temporal Patterns

- Wi-Fi Dataset provided by CISCO of a large European city
- 1.4 million devices, 91 days
- Evaluation using BLIPs, 1 BLIP per day



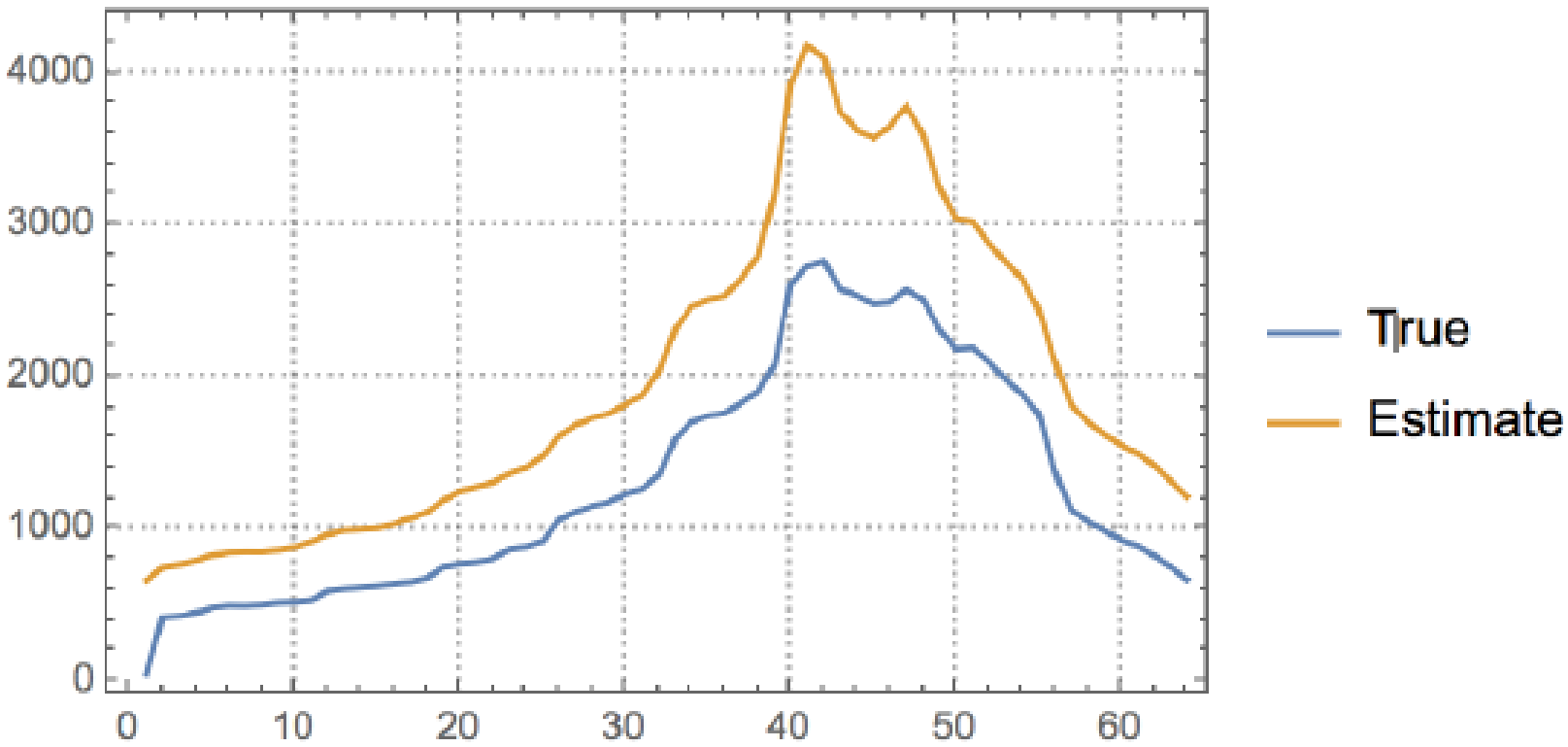
Spatial Patterns

- Top-10 origin-destination pair
- F1 score is 1 when two sets are identical and 0 if they share no elements at all



Temporal patterns (World cup dataset)

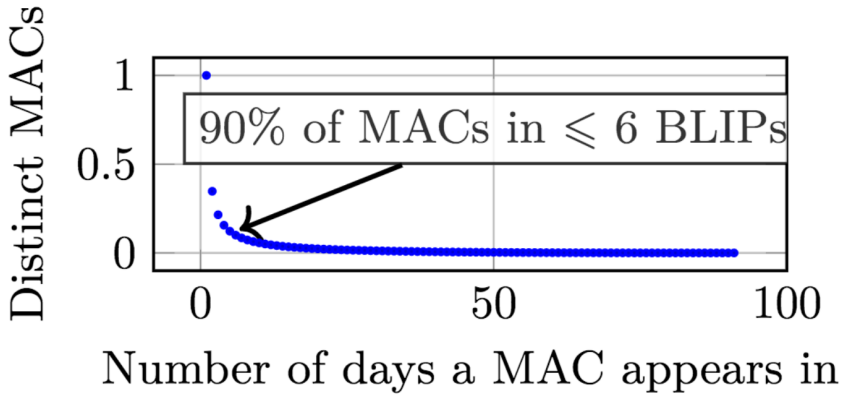
- HTTP request dataset for the FIFA World Cup 1998 website.
- 2.8 million unique IPs, 88 days.
- Evaluation using BLIPs, 1 BLIP per day ($\epsilon = 3$; $m = 2^{18}$)
- Estimating the **intersection of a rolling window of 30 days**



Managing the privacy budget

- **Fundamental issue of a privacy budget**: the more a user appears in several BLIPs, the more his privacy budget is impacted $\Rightarrow$ increase of risk of re-identification for a user.
- In practice, more than 90% of users do not appear in more than 6 BLIPs in the CISCO dataset
- **How to mitigate the impact**:

- Could change spatial or temporal granularity (make it more coarse)
- Regular change of hash functions (prevent inferences between BLIPs based on different hash functions) – not a silver bullet



Conclusion

Conclusion and Future Work

- Privacy-friendly wifi analytics: accurate patterns + privcay of individuals
- Pan-privacy: Privacy is preserved even if attacker gains full access to stored data
- BLIPs: Versatile building block for set operations
- We provide error bounds which can be of independent interest for **analysis of hashing collisions**
- Promising experimental evaluations
- Challenge: parameter tuning trade-off (ε, Bloom filter size)
  - Cardinalities are not known in advance
- Future work: Designing practical inference attacks
- Future work: More complex physical analysis tasks, e.g. traffic forecast, anomaly detection, point-to-point travel time, or urban network characterization

Thank You!