Private Attribution

PEARG, IETF 116
Martin Thomson
Trains
Identifiers

Access cards or credit cards provide a unique user/traveller identifier.

Travellers tap the card to enter and exit the system.

The distance between entry and exit can determine the fare due.

Designing a privacy-preserving system for charging fares based on route is a homework exercise.
Train Tracking

Many subway systems use tracking to monitor usage.

Traveller identifiers are logged on entry and exit:
- Cash payments or entry-only tracking (as in NYC, right) provide less information.

Queries of logs can reveal system utilization and can inform capacity planning.
Logs are a Privacy Risk

Logs contain extensive records on the movements of people
Each entry includes a time, a location, and an identifier
...and maybe more
Pseudonymous identifiers provide no meaningful privacy protection
Requirements

Be able to collected *aggregated* information about journeys

Protect details of individual user journeys
You hand the packets to the attacker to deliver

Issue people sealed tokens on entry

...then collect them at the exit

Entered at A

Exited at B
Privacy Mechanisms

Tokens need to be anonymous (or maybe really low entropy)
...or the token is just another identifier

Tokens need to be authenticated
...or people can lie

Opening tokens needs to be delayed
...or the timing reveals who it refers to

Random delays and anonymizing proxies might work

An aggregation system (PPM WG) can be faster
Token-Based Design Properties

Tokens are ephemeral
They are returned at the exit and only apply for that trip

Users carry tokens from the entrance to the exit

The information that a token provides is limited
This is generally good for privacy, with some caveats
...but this is inherently inflexible

Aggregation can help some of the worse aspects
Delays
Unknown anonymity set size
Trains | Advertising
Attribution

Attribution informs just about every aspect of advertising
Placement
Creatives
How much to spend
Attribution measures events that occur in different contexts to the same person

“How many people saw the ad then came to the show?”
Attribution is More Complex

Entering
- Happens once per trip

Exiting
- Happens once per trip

Showing ads
- ...or clicking ads
- ...or decided not to show an ad
- Happens 0..n times

Purchasing the product
- ...or just visiting the site
- ...or any outcome
- Happens 0..n times

Contextual data is irrelevant

Context is **everything**
Interoperable Private Attribution

People all have a secret identifier

Sites can ask for an encrypted copy of the identifier

Sites gather identifiers in logs along with contextual info

Sites cannot correlate user activity without help
IPA: Attribution in MPC

A site gathers events from multiple sites and uses contextual data to formulate a query.

MPC decrypts identifiers and performs attribution.

The result is aggregated results.
Multi-party computation can perform any computation
... without revealing individual inputs
All you need is additions and multiplications
... and money: complex computations can be very expensive
IPA uses a three-party, honest-majority MPC
... replicated secret sharing provides performance
... and almost information theoretic security guarantees
IPA is mostly generic MPC
Sorting groups inputs by the (hidden) identifier
Attribution is computed over adjacent inputs
IPA uses (ε, δ)-differential privacy to hide individual contributions.

Sites get a query budget that renews each epoch/week.

Privacy loss is bounded by time and number of sites involved:
- Each site has their own budget.
- Budgets are renewed weekly.
- Goal is to limit privacy loss rate.

Each query of the MPC uses up budget:
- Sites trade off noise with the number of queries.
- Values for ε and δ not decided.
Sensitivity Capping for DP

Encrypted identifiers are bound to
- The site that requested them
- The epoch/week they are requested
- The type of event: source (ad) or trigger (purchase)

Sites commit to using a single MPC (3 nodes)

Two types of query: source and trigger
- Source queries can only contain source events from one site
- Trigger queries can only contain trigger events from one site
- That one site expends its budget to make a query
- Site budgets are split evenly between the two types of query
IPA: Advantages and Challenges

IPA offers more flexibility for advertisers than alternatives
- Contextual information can be selected at query time
- Less need for special fraud prevention mechanisms

Flexibility might hurt accountability
- DP provides bounds on privacy loss, but no one understands DP
- The content of queries cannot be easily inspected and understood

MPC performance is a challenge
- Current implementation has plausible costs at small scale
- Scaling to meet needs of large advertising businesses is hard
Status

IPA is still active research
  Feasibility largely established
  Finer details of algorithms still being worked out
  Meta and partners are running trials

Ongoing work in the PATCG and PATWG in the W3C
  Other proposals are also being considered
  Protocols will likely go to IETF PPM WG