Minimal Viable Functionality for Measurement

Mariana Raykova
Google

Why common API?

A common API that is supported across different browsers and apps is an important goal that will most effectively meet the web ecosystem needs

- **There seems to be agreement**

Why common API even if restricted?

An API with the above properties is important even if it **does not meet everybody's complete set of requirements** (*different parties may have different extensions simultaneously*) :

Pros:

- Implementing and deploying something simpler will surface system challenges, which can be resolved before building complexity
- This is a first step, which can be extended later

Cons

- Modifications to the V1 may be hard and take a long time (we should think how it will be easy to iterate)
- Different parties may have different notions of MVP - we need agreemnt

Functionality Requirements

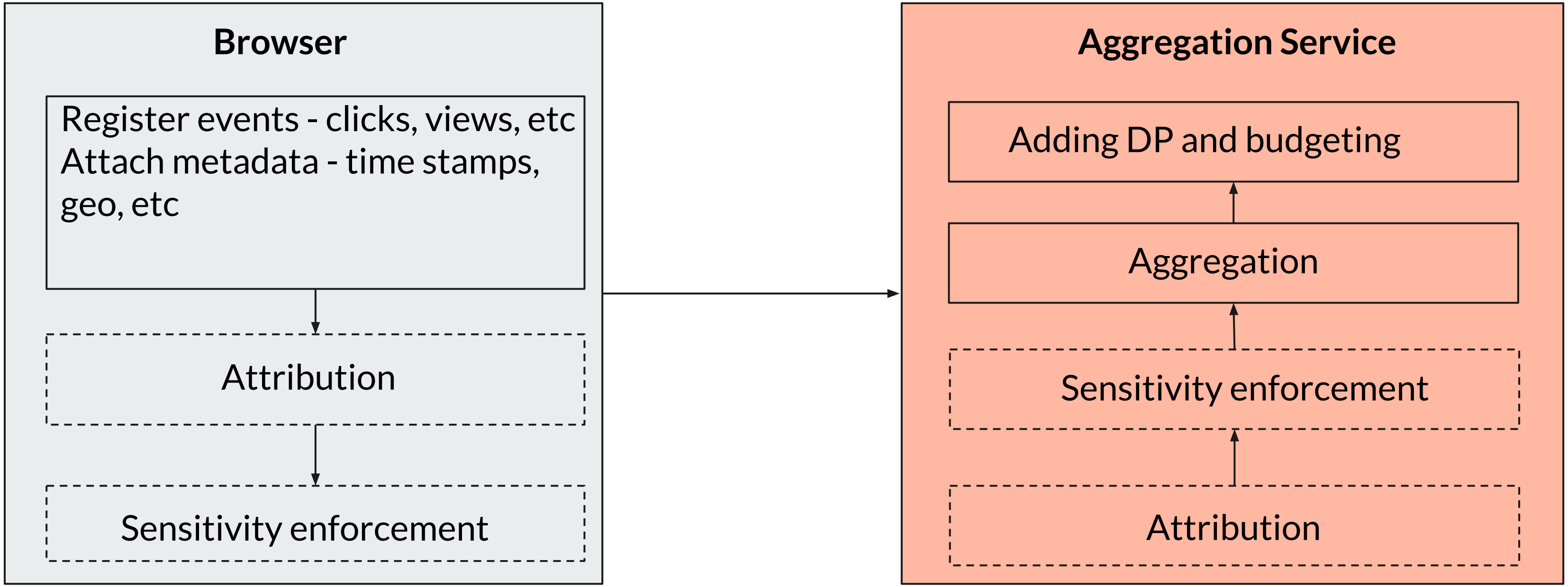
Aggregation vs Machine Learning

- Aggregation is useful on its own - count and value, building block for some simplifies ML
- Machine learning functionality will be needed to address all use cases eventually

Attribution on Device or off Device

- On device reduces complexity and security perimeter that needs to be protected with privacy preserving computation
- Off device offers cross-device attribution which is valuable in many settings

Strawman proposals - aggregation only



Browser

Register events - clicks, views, etc
Attach metadata - time stamps, geo, etc

Attribution

Sensitivity enforcement

Aggregation Service

Adding DP and budgeting

Aggregation

Sensitivity enforcement

Attribution

Most basic strawman functionality

Functionality:

- Only aggregation with DP
  - fixed number of aggregation buckets on the order of 2^{20} - 2^{32}
  - Sensitivity bound is fixed and applied on device
  - Adaptive epsilon per query
- Attribution on device

In the context of existing API proposals - differences:

- Chrome Attribution API - many more aggregation buckets 2^{128}
- Apple PCM - only 8 buckets and no DP
- Mozilla-Meta IPA - attribution off device on the servers
- Microsoft measurement proposal - more flexible aggregation buckets

Missing functionality

- Adaptively defined aggregation buckets at time of query
  - Importance - the aggregation granularity can be informed by different queries, e.g. for campaigns with many counts, consider finer-grained aggregation with attributes
- Different sensitivity bound per query
  - Different sensitivity bounds are optimal for different functionalities
    - Criteo challenge - logistic regression training: L2 bound sqrt(190), L1 bound 190
    - Count and value aggregates - L1 bound better result if the mass is on one key

Strawman functionality - adaptive query and DP

Functionality:

- Only aggregation with adaptive queries and dynamic DP budget per query
  - A query is specified by a set of **bucketing functions F_1,..., ,F_n** (chosen together)**, a sensitivity bound (e.g. L_0, L_1, L_{inf}) for the record contributions and an epsilon value**
  - The aggregation uses the following buckets: each record in mapped to a bucket F(record **impression data**)
  - The aggregators enforce the sensitivity bound for each record during aggregation
  - DP noise is added to the aggregated buckets according to the specified epsilon
- Attribution on device

Optionality for Extensions

- “High-level architecture of the API” allows flexibility for future use cases - how do we formalize this requirement?
    - Simple ML like logistic regression supported through the flexible aggregation
    - If we have an architecture where we are sending secret shared data, changes in functionality affect mostly the aggregation service side.
    - We should aim to minimize what goes out of the browser, but this can be changed in this architecture

Threat model

- Initial Proposal
  - Two parties - minimizes trust assumption
  - Malicious for privacy but semi-honest for correctness
    - Relaxation allow limited DP leakage to servers in the form of DP counts associated that cannot be linked to meaningful other impression, conversion data
  - Client input authentication - optional, can have approximation with trust tokens
- Cost - come up with a measure of acceptable cost