Designing an Open-Source Platform for Differentially Private Analytics That Is Usable, Scalable, and Extensible

Michael Hay
Tumult Labs & Colgate University
@michaelghay
● A system for safely releasing aggregate information from sensitive datasets

● Supports standard transformations (filter, join, map, ...) and aggregations (count, average, quantiles, ...)

● Currently used by:
An easy-to-use API capable of powering real-world use cases...
An easy-to-use API capable of powering real-world use cases...

... with provable guarantees of differential privacy...
An easy-to-use API capable of powering real-world use cases...

... with provable guarantees of differential privacy...

... that is extensible...
An easy-to-use API capable of powering real-world use cases...

... with provable guarantees of differential privacy...

... that is extensible...

... and can scale to arbitrarily-sized datasets.
Tumult Analytics

Intended for data scientists

- DP expertise *not* required
- Python interface similar to pandas/spark
session = Session.from_dataframe(
    dataframe=private_data,
    source_id="my_data",
    PureDPBudget(1.5)
)

query = (  
    QueryBuilder("my_data")  
       .filter("age > 42")  
       .groupby(zip_codes)  
       .median("income", low=0, high=10**6)
)

result = session.evaluate(
    query,  
    PureDPBudget(0.2)
)

print(session.remaining_privacy_budget())
# prints PureDPBudget(1.3)
User need

Maximal accuracy

Data-adaptive workflows

“Complex” data transformations

Features required

- tight privacy loss accounting
- zero-concentrated DP
- generalized parallel composition
- generalized stability calculus

- interactive mechanisms
- interactive parallel composition

- Flatmaps, joins, count distinct,
  Bounding user contributions

“Complex” data transformations
An easy-to-use API capable of powering real-world use cases...

... with provable guarantees of differential privacy...

... that is extensible...

... and can scale to arbitrarily-sized datasets.
Tumult Core

A collection of composable components: transformations and measurements.

Privacy properties
- Input domain
- Input metric
- Output domain
- Output metric
- Stability

Privacy properties
- Input domain
- Input metric
- Output measure
- Privacy loss
Tumult Core

A collection of composable components: transformations and measurements.

Privacy properties are derived inductively.

Combinators create new components from existing ones.
Tumult Core

A collection of composable components: transformations and measurements.

Combinators create new components from existing ones.

Includes floating-point safe implementations of critical mechanisms [Haney et al. TPDP 2022]
**Tumult Core**

A collection of composable components: transformations and measurements.

Core enables creation of complex algorithms from building blocks.

Everything carries an explicit, inspectable privacy guarantee.

Architecture is modular, extensible.
An easy-to-use API capable of powering real-world use cases...

... with provable guarantees of differential privacy...

... that is extensible...

... and can scale to arbitrarily-sized datasets.
Scalability Experiments

Simulations on synthetic dataset

Input rows: age, sex, detailed race, geography

Output: histograms (age x sex) for each race group at varying levels of geography

<table>
<thead>
<tr>
<th>Input size (rows)</th>
<th>Input size (GB)</th>
<th>Output size (rows)</th>
<th>Output size (GB)</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>300 million</td>
<td>41.3</td>
<td>62 million</td>
<td>1.30</td>
<td>7h 30m</td>
</tr>
<tr>
<td>600 million</td>
<td>82.7</td>
<td>64 million</td>
<td>1.37</td>
<td>14h 4m</td>
</tr>
<tr>
<td>900 million</td>
<td>115.0</td>
<td>65 million</td>
<td>1.39</td>
<td>20h 35m</td>
</tr>
</tbody>
</table>

Experiments were run on r4.8xlarge (1 master, 2 worker nodes)
An easy-to-use API capable of powering real-world use cases...

... with provable guarantees of differential privacy...

... that is extensible...

... and can scale to arbitrarily-sized datasets.
An easy-to-use API capable of powering real-world use cases and can scale to arbitrarily-sized datasets with provable guarantees of differential privacy.

🎉 Open-source release: today 🎉

[gitlab.com/tumult-labs/analytics](gitlab.com/tumult-labs/analytics)
[gitlab.com/tumult-labs/core](gitlab.com/tumult-labs/core)

We'd love your feedback!
(See slack for details)
Thank you!

Michael Hay
michael@tmlt.io
@michaelghay

gitlab.com/tumult-labs/analytics
gitlab.com/tumult-labs/core

tmlt.io/connect
tmlt.io/careers
@TumultLabs