Summary Report Optimization in ARA (Aggregate API)

WICG Review on Blog Post

aksu@google.com,

10 December 2023

Google

Outline

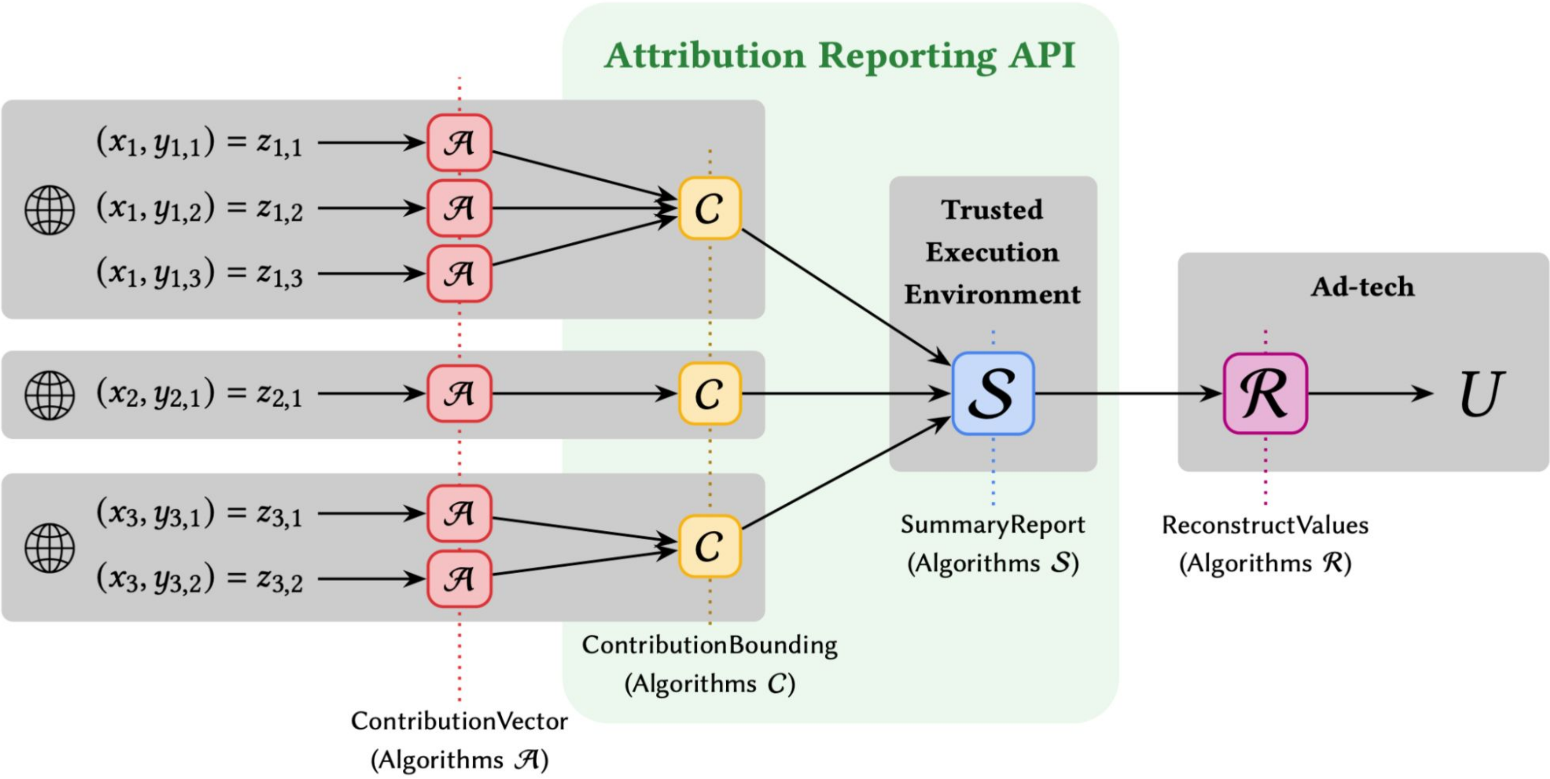
- Context & Problem
  - ARA Summary Reports
  - Maximizing the utility of summary reports
- Approach:
  - Mathematical model
  - Error metric
  - Optimization
- Synthetic Data
- Experimental Evaluations & Results
- Conclusions

ARA Summary Reports

	Impression features x			Conversion features y	
	Impression ID	Campaign	City	#items	value ($)
$z_1 \rightarrow$	123	Thanksgiving	New York	3	21
$z_2 \rightarrow$	123	Thanksgiving	New York	1	5
$z_3 \rightarrow$	456	Thanksgiving	Boston	1	99
$z_4 \rightarrow$	123	Thanksgiving	New York	2	23
$z_5 \rightarrow$	101	Christmas	Boston	2	50
$z_6 \rightarrow$	789	Christmas	New York	3	15
$z_7 \rightarrow$	101	Christmas	Boston	1	5
...	...	...	...	...	...

Impression and conversion feature logs for a fictional gift shop called Du & Penc

Mathematical Model



Google

Error Metrics

We have chosen:

- τ-truncated root mean square relative error

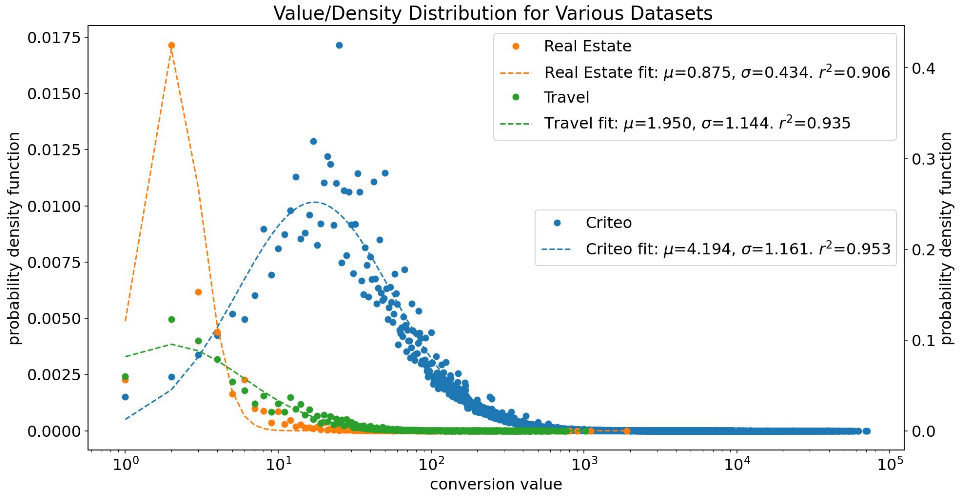
$$RMSRE_T(noise,\ true) = \sqrt{E([\frac{|noise|}{\max\{T, true\}}]^2)}$$

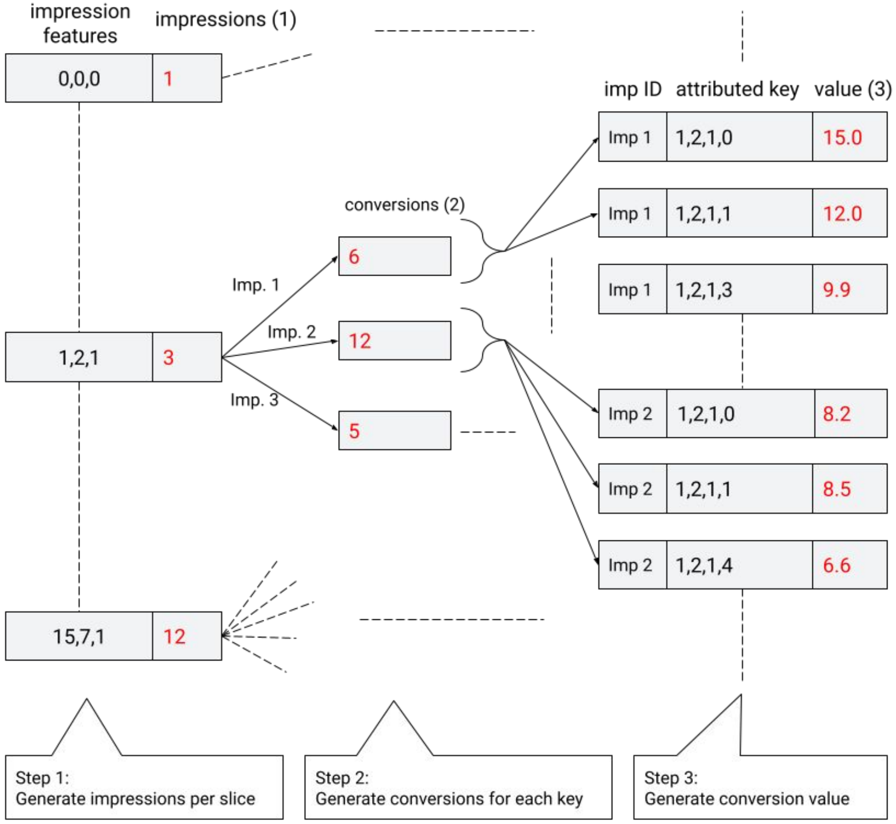
Optimization

- The problem
  - maximize the utility of summary reports
- Reduce to optimization problem
  - optimize utility as measured by RMSREτ,
    - capping parameter, *C*,
    - privacy budget, α, for each slice
- Observations
  - RMSREτ = the bias from clipping and the variance of the noise distribution.
  - for for a fixed privacy budget, α, or a capping parameter, C,
    - finding the optimal params is convex
  - for joint variables (C, α) it becomes non-convex
  - off-the-shelf optimizers works

Synthetic Data

- Generated synthetic data using
  - power law,
  - Poisson
  - log normal distributions





(1) Impressions drawn from power-law distribution.
(2) Conversions drawn from Poisson distribution.
(3) Conversion value drawn from log-normal distribution.

Experimental Evaluation

Datasets:

- Criteo: 15M clicks
- Real Estate: 100K conversions
- Travel: 30K conversions
- 3 Synthetic ones

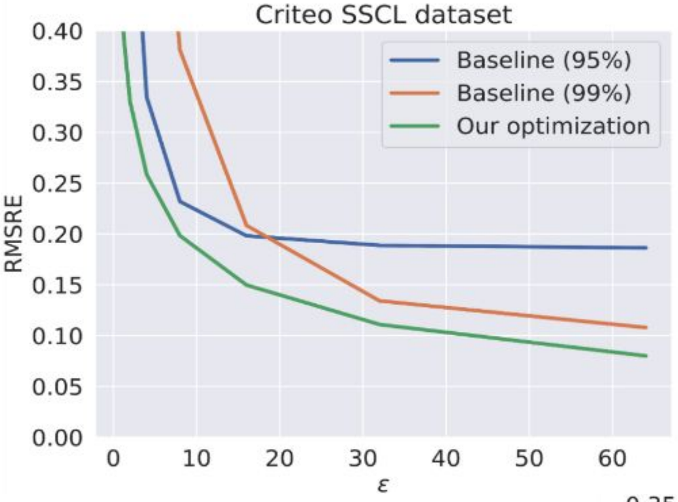
Dataset Partitioning:

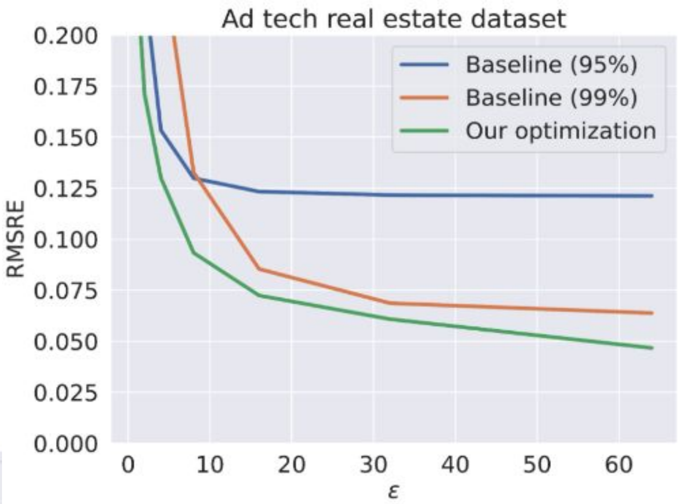
- Training set: Choose budgets, thresholds, and limits
- Test set: Evaluate errors

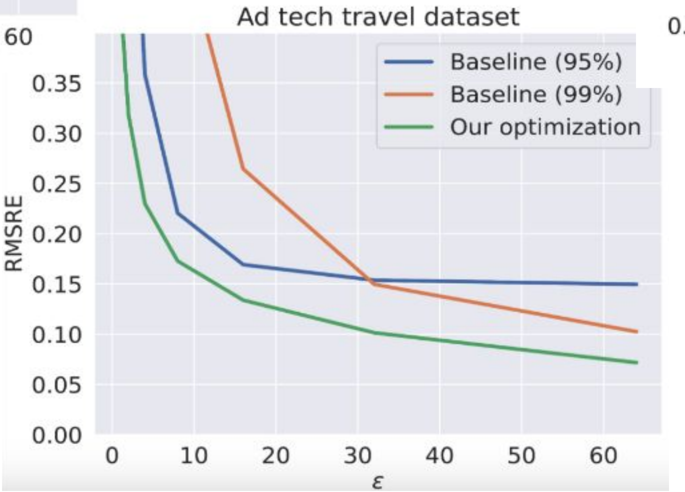
Error Metric: RMSRE$\boldsymbol{\tau}$ τ:

- Chosen to be 5 times the median value on training data
- Invariant to data rescaling
- Allows combining errors across different scales

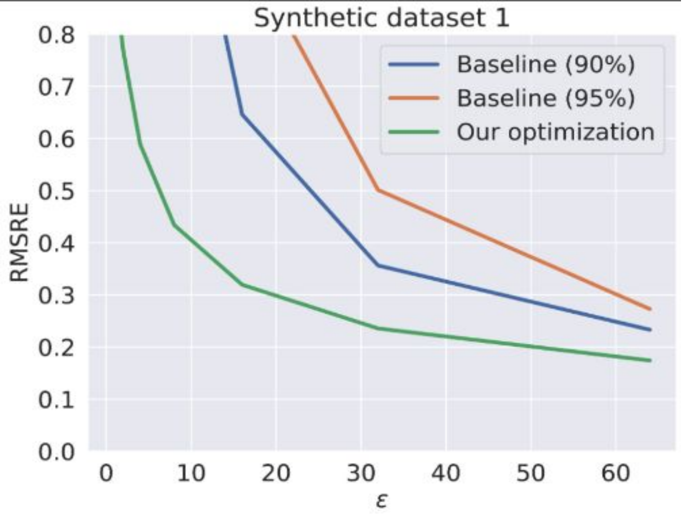
Results

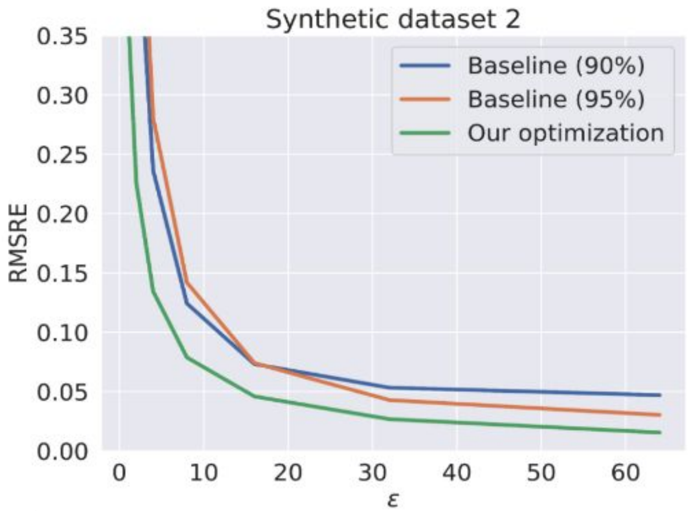


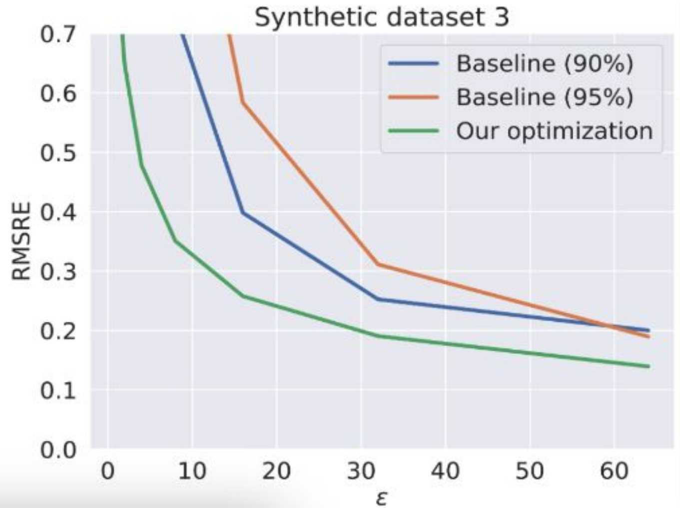




Baselines that use a fixed quantile for the clipping threshold and split the contribution budget equally among the queries.


Synthetic dataset 1
RMSRE
ε
Baseline (90%)
Baseline (95%)
Our optimization


Synthetic dataset 2
RMSRE
ε
Baseline (90%)
Baseline (95%)
Our optimization


Synthetic dataset 3
RMSRE
ε
Baseline (90%)
Baseline (95%)
Our optimization

Conclusion

- Leverages historical data
    - bound and scale the contributions of future data
    - use synthetic data
- Paper provide `Generalization bounds`
    - we're not overfitting to the historical data
- Blog post, the paper and accompanying code are public.