Private Cross-Media Reach & Frequency Estimator Evaluation

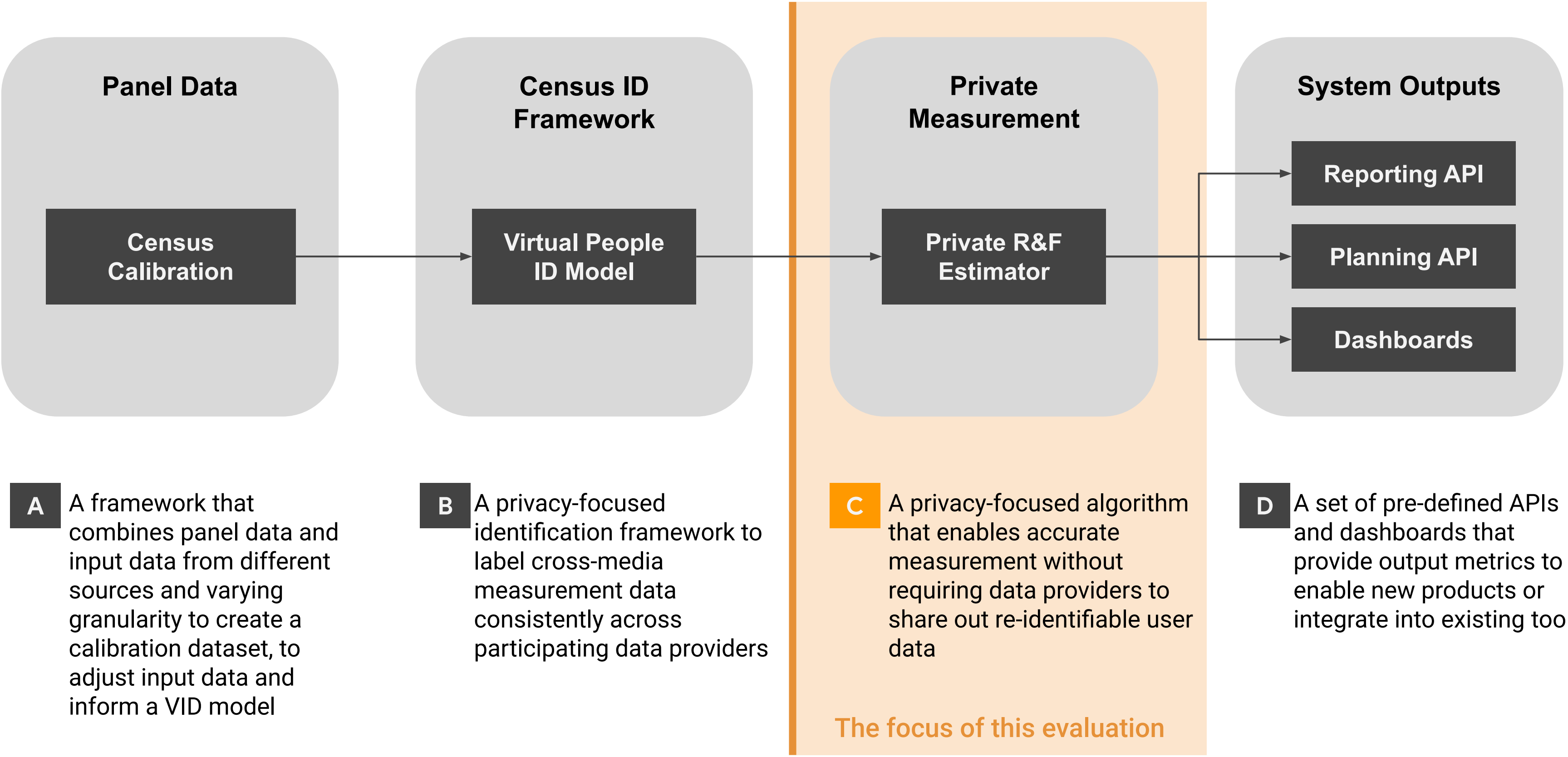
Preliminary Results & Recommendations

Executive Summary

- We recommend a privacy preserving reach and frequency estimation algorithm, called **Liquid Legions**, that uses a secure multiparty computation (MPC) protocol for combining inputs across multiple impression data providers.
- The recommended algorithm excels in
  - **Accuracy** in estimating cross-media reach and frequency
  - **Strong Privacy protection** against user re-identification
  - **Scalability** where its merits are consistent across cross-media campaigns of different sizes, different numbers of data providers, and different sizes of exposure overlaps.

Four Key Components of Cross-media Measurements

Recap from WFA Cross-Media Blueprint



A — A framework that combines panel data and input data from different sources and varying granularity to create a calibration dataset, to adjust input data and inform a VID model

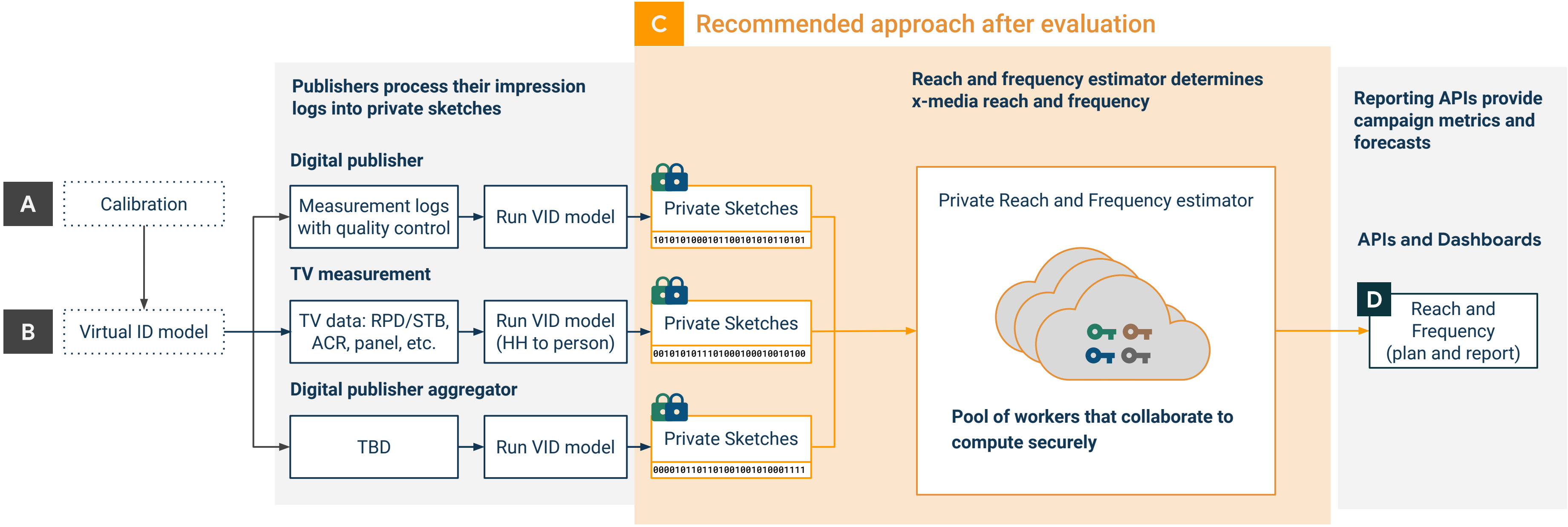
B — A privacy-focused identification framework to label cross-media measurement data consistently across participating data providers

C — A privacy-focused algorithm that enables accurate measurement without requiring data providers to share out re-identifiable user data

The focus of this evaluation

D — A set of pre-defined APIs and dashboards that provide output metrics to enable new products or integrate into existing tools

How Does Private RF Estimator Fit In The Measurement Phase



C Recommended approach after evaluation

A Calibration

B Virtual ID model

Publishers process their impression logs into private sketches

Digital publisher

Measurement logs with quality control

Run VID model

Private Sketches

1010101000101100101010110101

TV measurement

TV data: RPD/STB, ACR, panel, etc.

Run VID model (HH to person)

Private Sketches

0010101011101000100010010100

Digital publisher aggregator

TBD

Run VID model

Private Sketches

0000101101101001001010001111

Reach and frequency estimator determines x-media reach and frequency

Private Reach and Frequency estimator

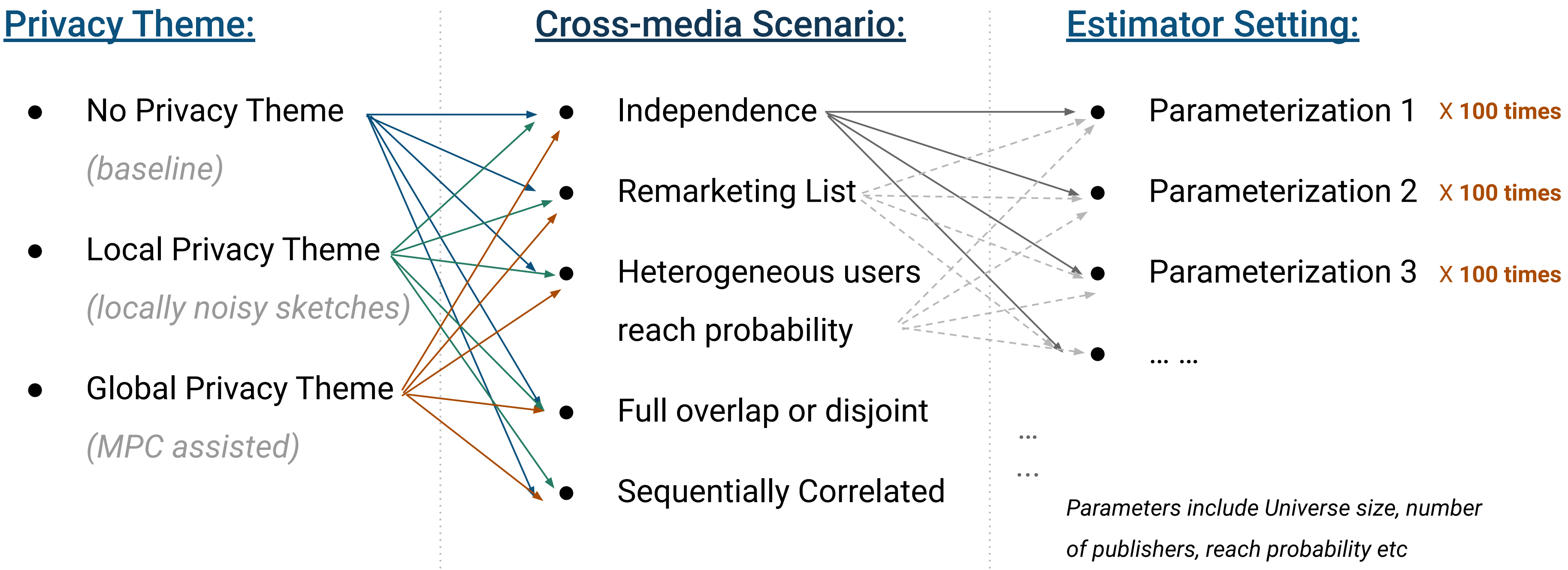
Pool of workers that collaborate to compute securely

Reporting APIs provide campaign metrics and forecasts

APIs and Dashboards

D Reach and Frequency (plan and report)

Theme, Scenario and Setting



Privacy Theme:

- No Privacy Theme

  (baseline)

- Local Privacy Theme

  (locally noisy sketches)

- Global Privacy Theme

  (MPC assisted)

Cross-media Scenario:

- Independence
- Remarketing List
- Heterogeneous users reach probability
- Full overlap or disjoint
- Sequentially Correlated

…

…

Estimator Setting:

- Parameterization 1 X 100 times
- Parameterization 2 X 100 times
- Parameterization 3 X 100 times
- … …

Parameters include Universe size, number of publishers, reach probability etc

Private RF Estimator Candidates: Two Groups

Group 1: Bloom filter variants

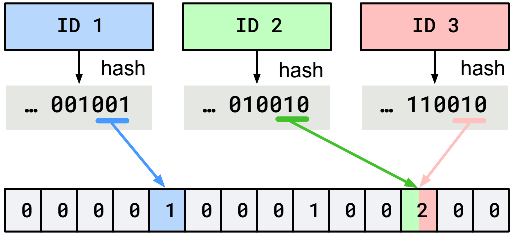
- Exponential Bloom Filters
- Logarithmic Bloom Filters
- Geometric Bloom filters
- Liquid Legions *(Exponentially Distributed Counting Bloom filter with Same Key Aggregator)*

Group 2: Non-Bloom-filter-based Private Sketches

- The Conditional Independence Model
- Vector of Counts
- Meta-VoC
- Hyper Log Log ++
- Stratified Sketches

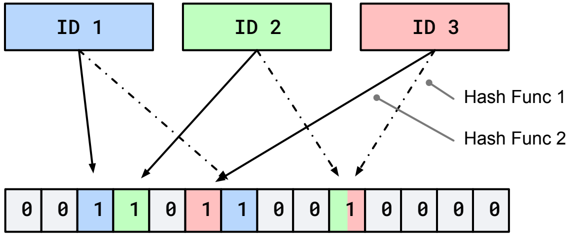
Illustrations of Different Private RF Estimator

1 Vector of Counts



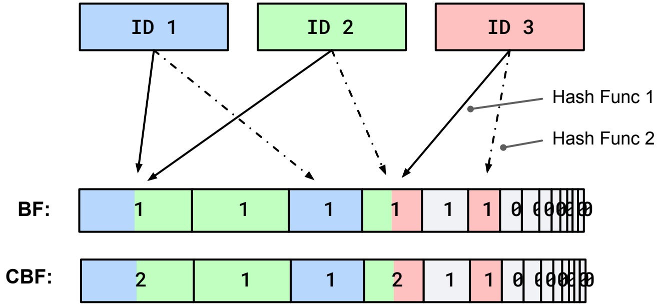
0 0 0 0 1 0 0 0 1 0 0 2 0 0

2 Uniform Bloom Filters



0 0 1 1 0 1 1 0 0 1 0 0 0 0

3 Any Distribution Bloom Filters and Counting Bloom Filters



BF: 1 1 1 1 1 1 0 0 0 0 0

CBF: 2 1 1 2 1 1 0 0 0 0 0

Evaluation Results Summary

No Privacy Theme Reach Results

Cross-media Reach Scenarios (95/5 criteria)	Meta VoC Uni	Meta VoC Exp	Exp BF 10	Geo BF	HLL++	Log BF	Cond. Ind.	VoC
Independence model	13.83	0	20	20	20	20	20	17.67
Remarketing List	17	0	20	20	20	20	1.83	18.83
Heterogeneous reach probability: Independent reach prob	14.33	0	20	20	20	20	20	19.17
Heterogeneous reach probability: Fully overlapped reach prob	10	0	20	20	20	20	5.83	10.33
Fully overlapped universe: num_sets-20	10	0	20	20	20	20	0	20
Fully overlapped universe: num_sets-random	8.33	0	20	20	20	20	0	8.33
Sequentially correlated order random correlated sets: one	10.31	0	20	20	20	20	0.06	10.22
Sequentially correlated order random correlated sets: all	9.72	0	20	20	20	20	1.31	11.83

No Privacy Theme Frequency Results

Cross-media Frequency Scenarios (95/5 criteria)	Liquid Legions (max freq = 15)	Liquid Legions (max freq = 5)	Strat VoC Clip (max freq = 15)	Strat VoC Clip (max freq = 5)
1.homogeneous-universe_size:200000-num_sets:10	10	10	4.25	6.08
2.heterogeneous-universe_size:200000-num_sets:10	10	10	2.67	3.83
3.publisher_constant_frequency-universe_size:200000-num_sets:10	10	10	10	10

No Privacy Theme Takeaways

- Without differential privacy noise, HLLs and Bloom filter-based techniques are capable of unioning any number of sets without any degradation in accuracy.
- The other methods do not perform as well and are impacted by the input data distribution (scenarios).
- When independence assumption is violated, several estimators from group 2 exhibit bias.

Local Privacy Theme Reach Results

Cross-media Reach Scenarios (95/10 criteria | epsilon = ln(3))	Meta VoC Uni.	Meta VoC Exp	Exp BF 10	Exp BF 2	Geo BF	Log BF	Cond. Ind.	VoC
1.independent-universe_size:1000000-small_set:10000	19.67	1	2.5	6.67	6.67	5.67	20	20
2.remarketing-remarketing_size:200000-universe_size:1000000	20	1	2.67	6.17	6.17	5.5	2.83	20
3a.exponential_bow-user_activity_association: independent-universe_size:1000000	19.67	1	2.33	6	6.67	5.5	20	20
3b.exponential_bow-user_activity_association: identical-universe_size:1000000	11.33	1	2.83	6.67	6.5	6	9	11.5
4a.fully_overlapped-universe_size:1000000-num_sets:20	20	0	2	4	4	3.5	0	20
4b.subset-universe_size:1000000-order:random	8.67	0	1.33	2.67	2.67	2.33	0	8.67
5a.sequentially_correlated-order:random-correlated_sets:one	13.44	0.67	2.89	6.56	6.39	5.81	0.08	13.08
5b.sequentially_correlated-order:random-correlated_sets:all	13.56	0.69	3.14	6.56	6.67	6.08	3.22	14.92

Local Privacy Theme Frequency Results

Cross-media Frequency Scenarios (95/10 criteria | epsilon = `ln(3)`)	Stratified_VoC-5 Clip	Stratified_VoC-15 Clip
1.homogeneous	2.92	1.58
2.heterogeneous	2.58	1.5
3.publisher_constant	6.75	2.25

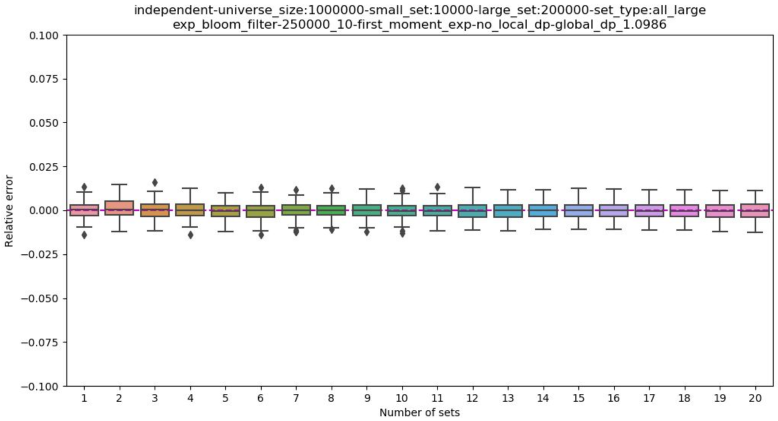
Local Privacy Theme Takeaways

- In this privacy theme all of the Bloom filter variants perform similarly, and were able to union at least four sets in all scenarios with 95/10 accuracy.
- Besides the noise level, another important factor that impacts Bloom filter accuracy in this theme is the size of the Bloom filter with respect to the size of the set to be measured.
- Conditional independence works well when there is actual independence and rather poorly otherwise. It is not a serious contender in this theme.
- Vector of counts performs very well across many scenarios, but as in the no privacy theme it exhibits significant bias when input data is correlated

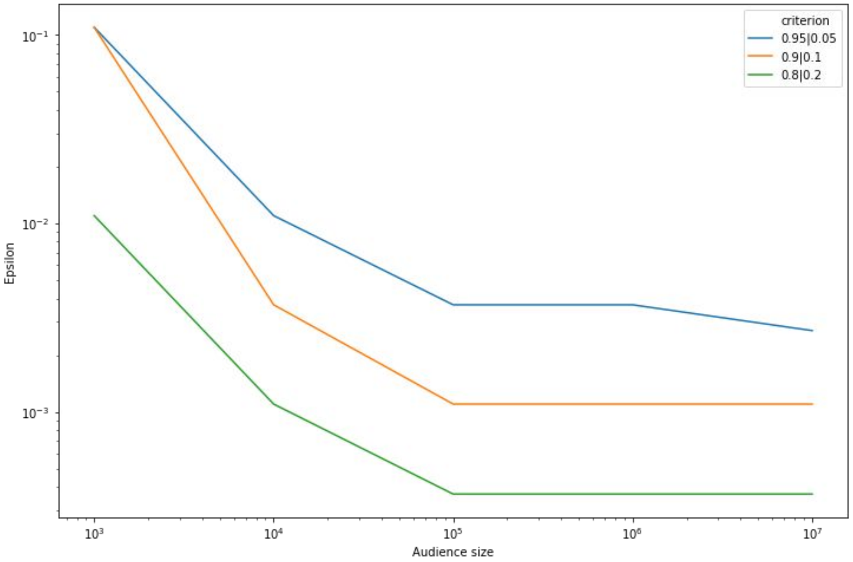
Global Privacy Theme Reach Results

Cross-media Reach Scenarios (95/5 criteria)	Exponential Bloom filter	Geometric Bloom filter	Log Bloom filter
1.independent-universe_size:1000000-small_set:10000	20	20	20
2.remarketing-remarketing_size:200000-universe_size:1000000	20	20	20
3a.exponential_bow-user_activity_association:independent-universe_size:1000000	20	20	20
3b.exponential_bow-user_activity_association:identical-universe_size:1000000	20	20	20
4a.fully_overlapped-universe_size:1000000-num_sets:20	20	20	20
4b.subset-universe_size:1000000-order:random	20	20	20
5a.sequentially_correlated-order:random-correlated_sets:one	20	20	20
5b.sequentially_correlated-order:random-correlated_sets:all	20	20	20

Scalability: Union cardinality error remains nearly constant as number of set increases.



Minimum epsilon stabilizes (y-axis) as estimable audience size (x-axis) increases.



Global Privacy Theme Frequency Results

Cross-media Frequency Scenarios (95/5 criteria | epsilon = `ln(3)`)	**Liquid Legions - 15**
1.homogeneous-universe_size:200000-num_sets:10	10
2.heterogeneous-universe_size:200000-num_sets:10	10
3.publisher_constant_frequency-universe_size:200000-num_sets:10	10

Global Privacy Theme Takeaways

- Only Bloom filter based estimators are considered because a MPC protocol is required for this theme.
- Bloom filter based estimators all accurately estimate a 20-set union and based on mathematical theory we expect no problem in scaling beyond 20 sets.
- With the MPC protocol, the accuracy of the RF estimates is not dependent on the number of sets to union, because the noise being added is the same.
- The only missing piece of data for global theme is the cost of running the MPC framework, which will definitely be more expensive than a system that uses techniques from the local privacy theme.

Private RF Estimator Recommendation

Overall Recommendation

- Pending the results of the MPC performance test we provisionally recommend to use the **Liquid Legions sketch along with a MPC protocol** for estimating cross media reach and frequency.
- There are several reasons for this recommendation:
    - The accuracy of MPC-based Liquid Legions sketch for reach measurement is invariant to the number of sets being unioned and the distribution of its inputs.
    - Its accuracy for frequency measurement is invariant to the maximum frequency being measured and the number of sets to be unioned.
    - Moreover, it is the only method that can measure frequency across multiple publishers with an acceptable max frequency. In short, it is highly accurate under all reach and frequency scenarios.

Local Implementation Implications

- Local data providers/publishers utilize the global Liquid Legion (LL) reference implementation to transform the VID-labeled cross-media impression data.
    - Develop local infrastructure to support LL sketches generation.
    - Test VID → Sketches global privacy theme transformation with synthetic data as part of the E2E test.
    - Engage local TV and digital publishers to standardize the sketches generation process.
- Develop local infrastructure to support and integrate with MPC-assisted sketches combination process.

References

- Github repository for the private reach & frequency estimator evaluation work
- Detailed paper on evaluation results
- Detailed paper on evaluation framework definition
- Detailed paper on different cardinality estimators, including Liquid Legions