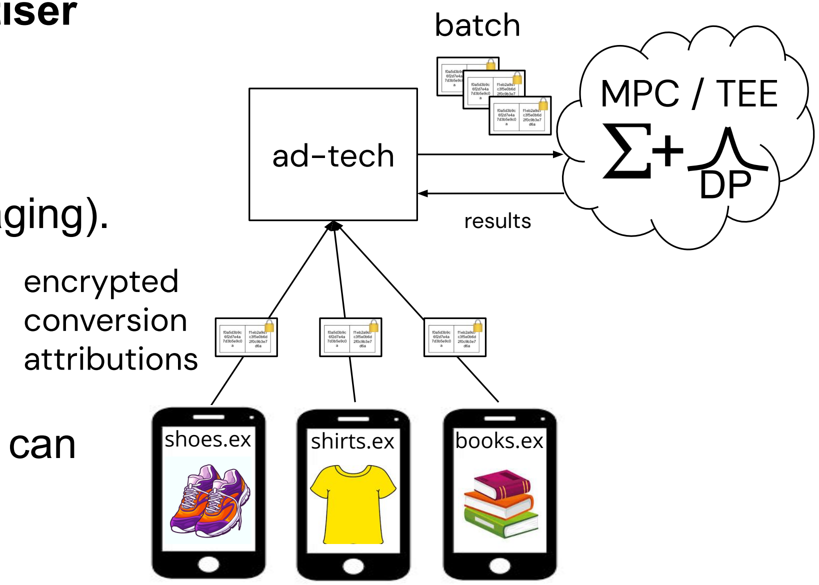
Evaluating Optimization Queries:
Methodology and Preliminary Results

Presenters:
Benjamin Case (Meta) and Roxana Geambasu (Columbia University, tmp. with Meta)

Work done by:
Mark Chen and Giorgio Cavicchioli (Columbia University),
advised by **Dr.** Pierre Tholoniat and the presenters

Attribution Level 1: Measurement Queries

- Current Attribution API supports **single-advertiser DP queries** (*measurement queries*).
- Advertisers can use these to evaluate **ad performance** (e.g., compare creatives, messaging).
- Works best for **large advertisers** with many conversions.
- Optimization support: **ad-tech intermediaries** can post-process per-advertiser DP outputs for **optimization purposes**, such as to learn ad-placement models, **without extra privacy loss.**



Our Goal:

Evaluate **optimization use cases** on:
single-advertiser queries (Attribution Level 1)
vs. cross-advertiser queries (envisioned for Level 2)

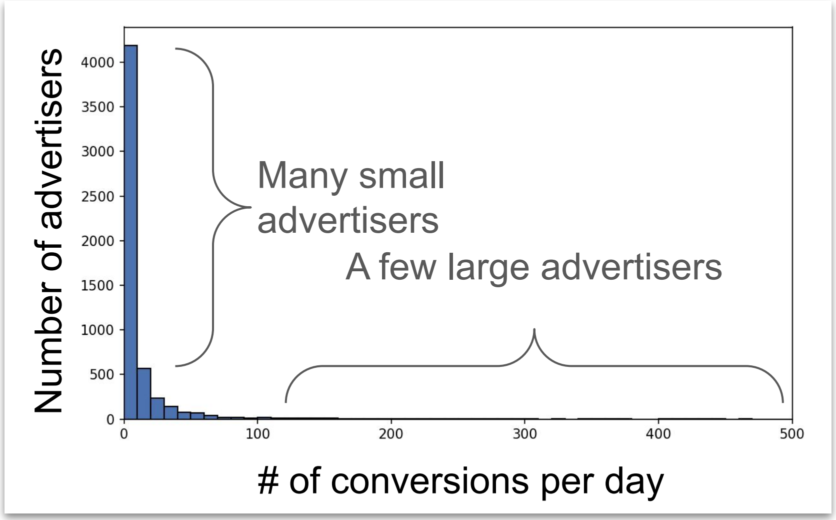
Non-goal today: ***how*** to support cross-advertiser queries in Attribution

Outline

- **Preliminary methodology**
- *Very* preliminary results
- Next steps
- Your feedback

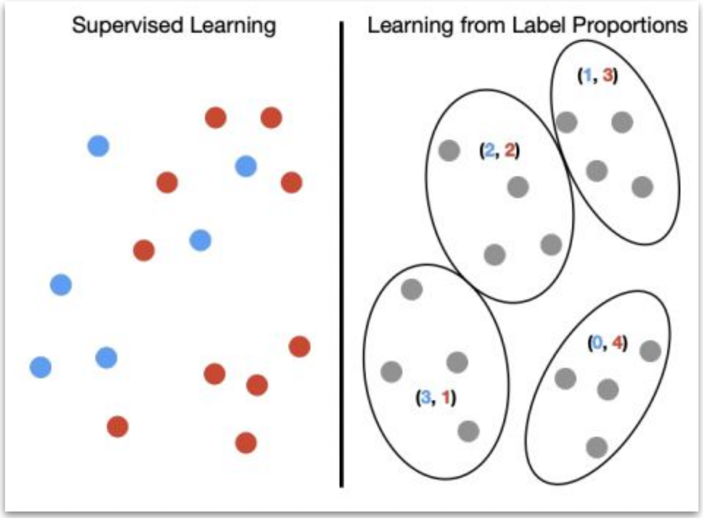
Methodology overview

- **Dataset**: Criteo dataset [SOL+25]
  - 100M impression entries: date, impression features, ad campaign, user ID, …, IS_CLICK.
  - 40k advertisers with 37% is_click "conversions."
- **Algorithm**: learning from label proportions (LLP) [BDG+25] -- well suited for Attribution.
- **Big/small advertisers:** extreme imbalance, 1% of advertisers account for 30% of impressions and 35% of conversions.
  - Small advertisers (<10 conversions or <30 impressions) account for 30% of the impressions and for XXX% of the conversions.
- **Task**: learn click-through rate (CTR) prediction model.



Background: LLP

- Trains on **"bags" of examples** with known **label proportions**, not individual labels.
- **Features** of each example are known; **labels** are hidden.
- Well-suited to the **Attribution API** (including Level 1), where aggregates provide **noisy label proportion estimates**.
- Mentioned in the Criteo paper as an example learning task for their dataset.

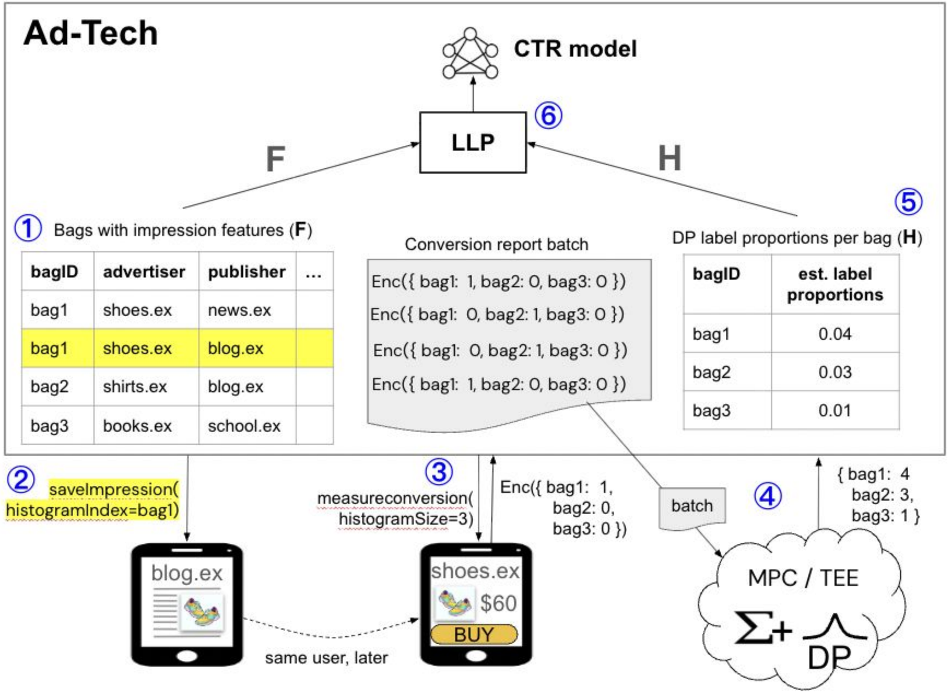


(fig. credit: [Busa-Fekete, et.al., 2023])

Loss: $\ell^c(h, z_j) = \frac{1}{k}\left(k\widetilde{\alpha}_j - \widetilde{\mathbb{E}}_j(h)\right)^2 G_j(h) + \left(\mathbb{E}[h(x)] - p\right)^2$

CTR training with LLP over Attribution API

- **SaveImpression():** ① Ad-tech assigns the impression to a "**bag**", saves impression features **F** in the backend, and ② invokes saveImpression(histogramIndex=bagID, ...).
- **MeasureConversion():** ③ Conversion produces a histogram mapping {bagID→{0,1} }.
- ④ **DP aggregation** over a batch of multiple reports yields {bagID→DP-estimated # of conversions}.
- ⑤ Estimate conversion rate per bag: **H** [bagID] = est_conversions / bag_size.
- ⑥ Feed **(F, H)** into **LLP** to train CTR model.



Bagging: Single- vs. Cross-Advertiser Queries

- **Single-advertiser:** Bags must be per advertiser (matching DP aggregates).
  - Large advertisers: can split into good-size bags (e.g., 30 impressions).
  - Small advertisers: too few impressions or conversions, leading to either too small bags or too small batches ⇒ high DP noise, weak signal.
- **Cross-advertiser:** Flexible bagging (random, per-advertiser, or mixed).
  - Small advertisers: can be grouped into good-size bags, and DP noise will be split among the entire cross-advertiser batch, impacting the small-advertiser signal less than with per-advertiser batch.
- We use:
  - Single-advertiser: impression's `bagID` is (advertiserID,random) (advertiser's bags).
  - Cross-advertiser: impression's `bagID` is random (all bags).

Research question:

For fixed privacy loss, does LLP lead to more accurate click prediction, **especially for small advertisers**, when trained over cross- vs. single-advertiser queries?

To test, we evaluate multiple settings

Baselines:

A. No LLP, no DP (Adam w/ fully-connected 2 layers + sigmoid, and **binary cross entropy** loss)
B. LLP, no DP, random bagging (same optimizer & model, but **LLP loss** ← same for all below)
C. LLP, no DP, per-advertiser bagging

Attribution API:

D. LLP, cross-advertiser DP query
E. LLP, single-advertiser DP query, drop too small bags & batches
F. LLP, single-advertiser DP query, keep too small bags & batches despite noise

Evaluate:

- CTR model accuracy on entire hold-out test set of impressions.
- CTR model accuracy on all vs. small-advertiser impressions ("small" are those that would be dropped from E).

Outline

- Preliminary methodology
- ***Very* preliminary results**
- Next steps
- Your feedback

Preliminary setup

- Results from notebook-based processing of Criteo dataset (no Attribution integration).
- We add DP noise to aggregates, but we don’t perform on-device budgeting, which means we effectively impose no per-site limits on privacy consumption.

- **Fixed params with limited or no tuning:**
  - **Data sample:** 400K impressions from a short time window of the Criteo dataset. Train:validation:test = 80:10:10, sets fixed upfront for each graph.
  - **Fixed bag size:** [20,30] impressions/bag (except for F, which admits smaller).
  - **Fixed DP noise:** Lap(0,b), b being the same for all DP lines in a graph.
  - **Privacy loss**: we can’t calculate individual privacy loss w/o Attribution integration, but we approximate the global privacy loss as ~1/b across all DP lines in a graph by assuming that sensitivity across all DP settings is the same (=1) and therefore ignoring the reality that a single user can participate with more than one conversion...
  - **“Small advertiser**”: <30 impressions or <30 conversions.

- Goal: early signal on hypothesis validity, but still ***very*** **preliminary results**.

CTR model ROC (preliminary)

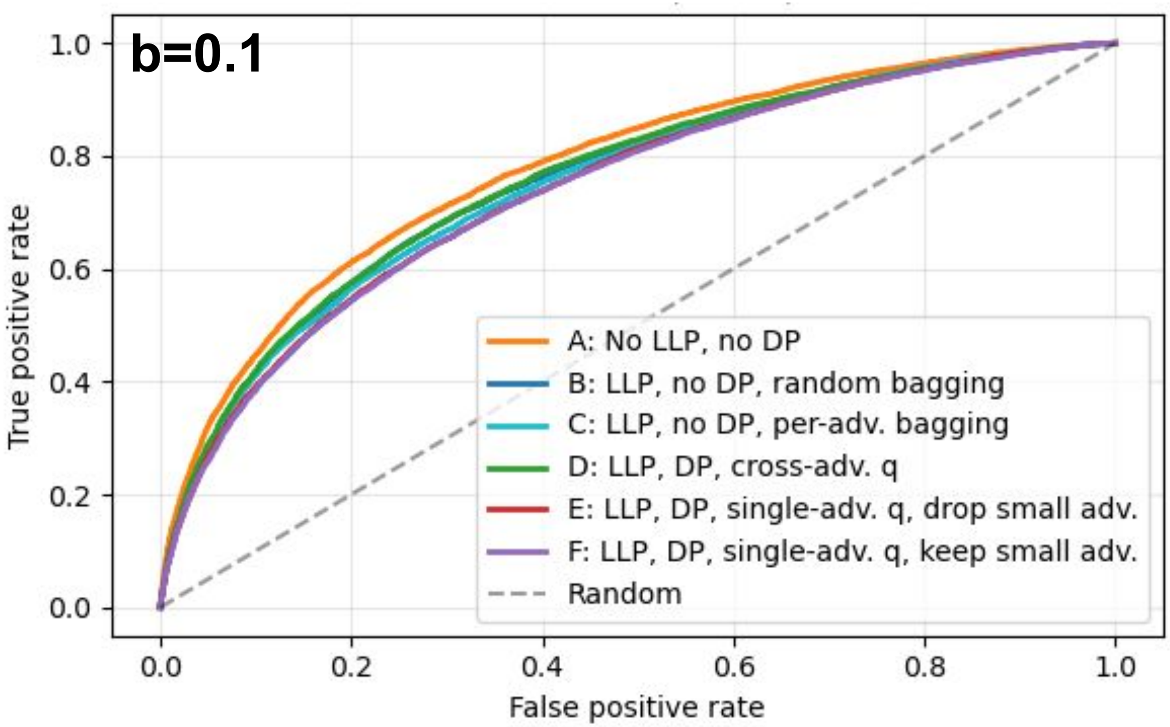
Params:

Bag sizes: A-E: ~30 impressions; F: ~30 or smaller.

DP noise: Lap(0, b): scale b fixed across DP queries in a graph.

V. approximate privacy loss for D-F: ε=~~1/b (if all sensitivities=1…).

Small advertiser: <30 impressions or conversions.



CTR model ROC (preliminary)

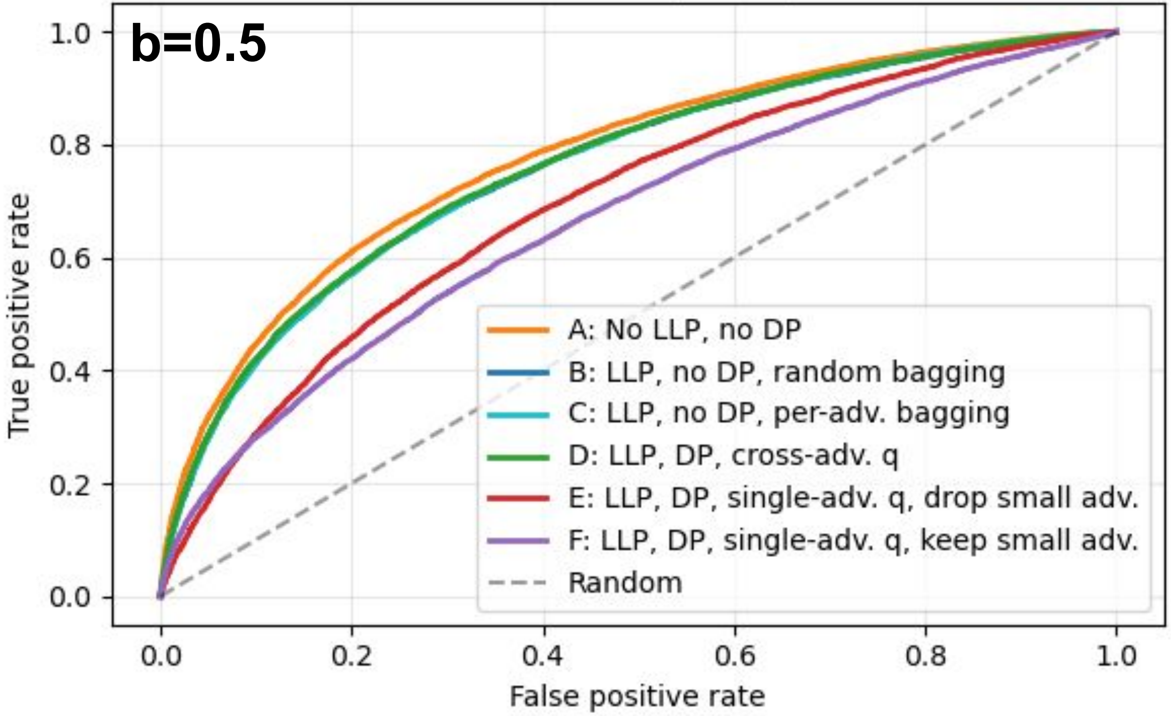
Params:

Bag sizes: A-E: ~30 impressions; F: ~30 or smaller.

DP noise: Lap(0, b): scale b fixed across DP queries in a graph.

V. approximate privacy loss for D-F: ε=~~1/b (if all sensitivities=1…).

Small advertiser: <30 impressions or conversions.



CTR model ROC (preliminary)

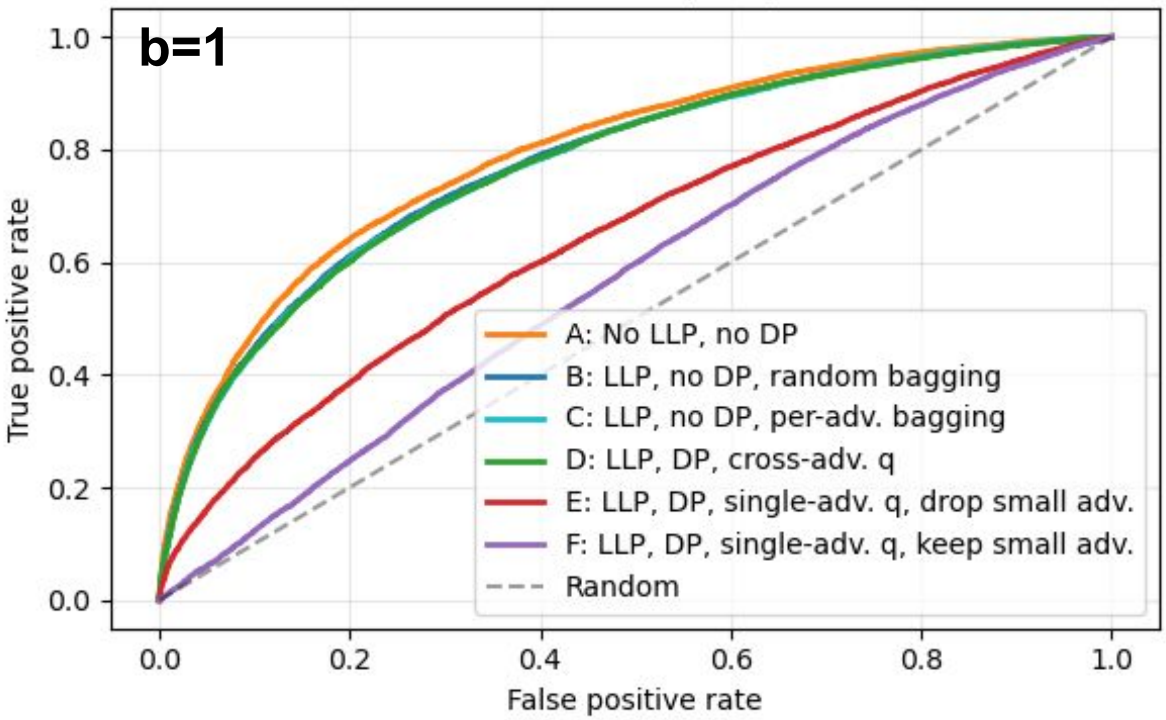
Params:

Bag sizes: A-E: ~30 impressions; F: ~30 or smaller.

DP noise: Lap(0, b): scale b fixed across DP queries in a graph.

V. approximate privacy loss for D-F: ε=~~1/b (if all sensitivities=1…).

Small advertiser: <30 impressions or conversions.



CTR model ROC (preliminary)

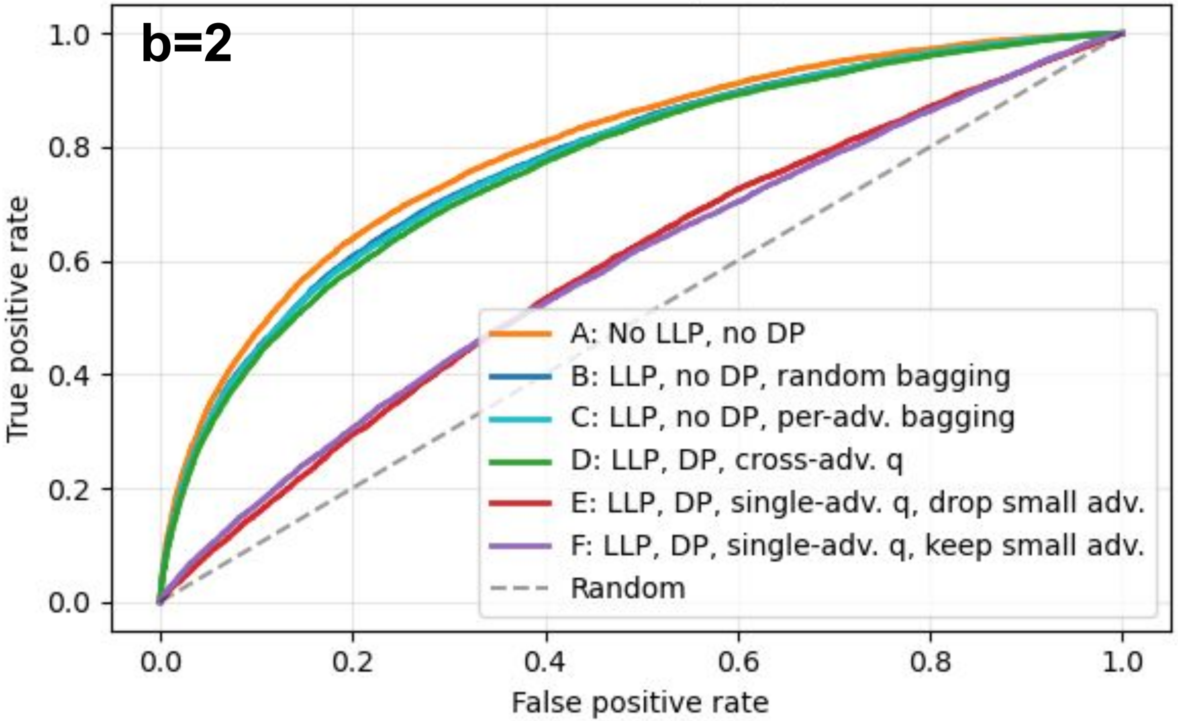
Params:

Bag sizes: A-E: ~30 impressions; F: ~30 or smaller.

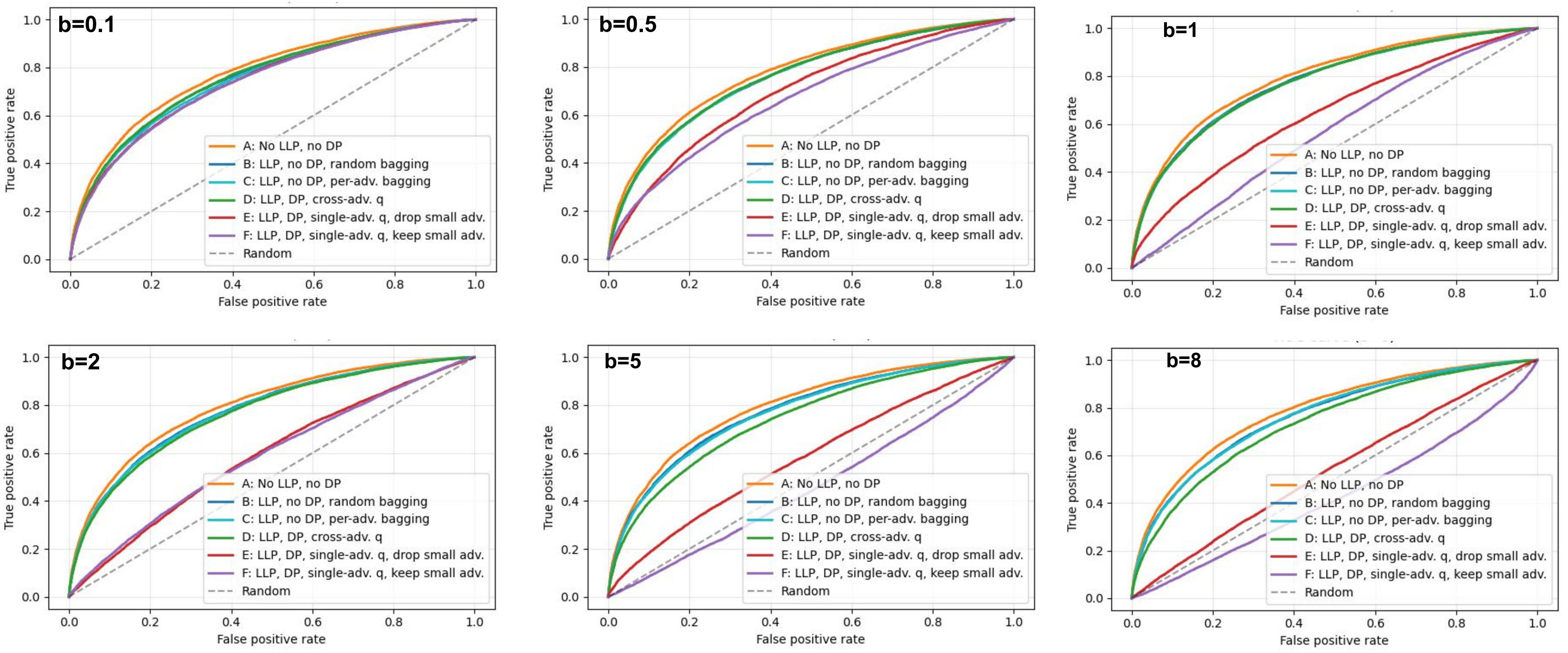
DP noise: Lap(0, b): scale b fixed across DP queries in a graph.

V. approximate privacy loss for D-F: ε=~~1/b (if all sensitivities=1…).

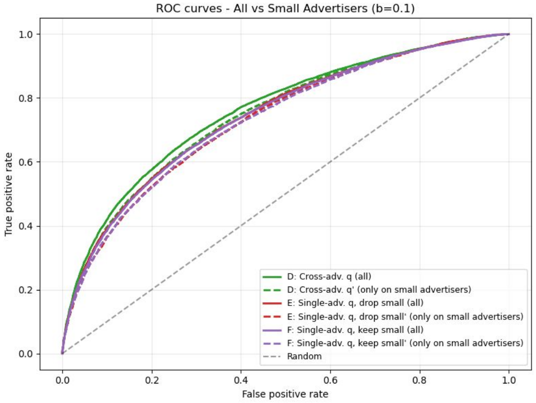
Small advertiser: <30 impressions or conversions.

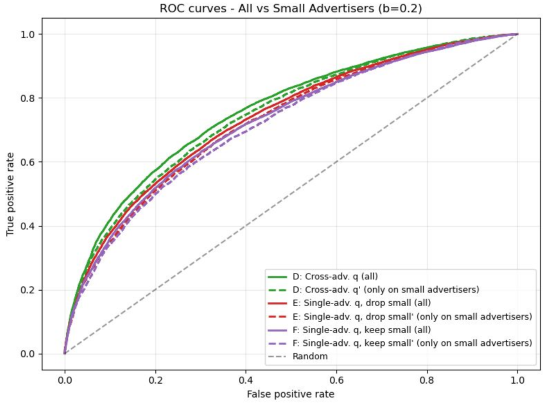


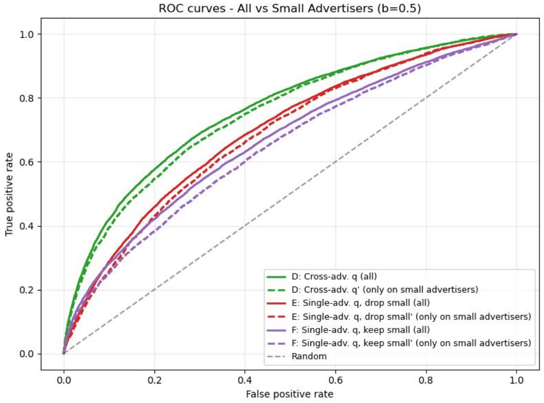
CTR model ROC (preliminary)

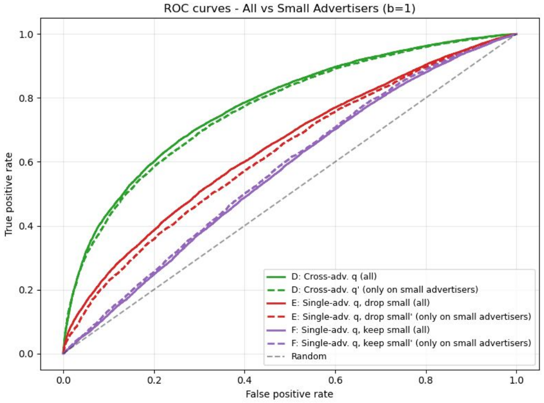


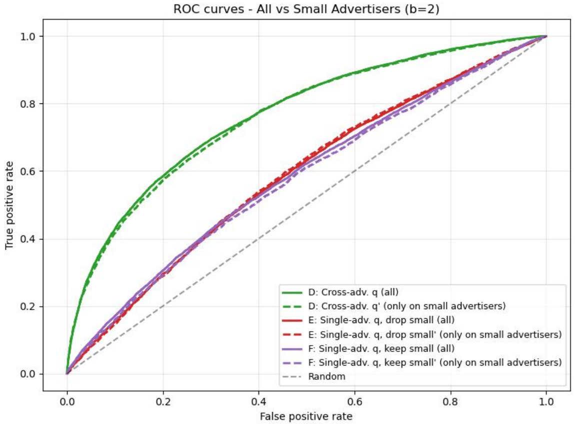
Utility for small advertisers (preliminary)

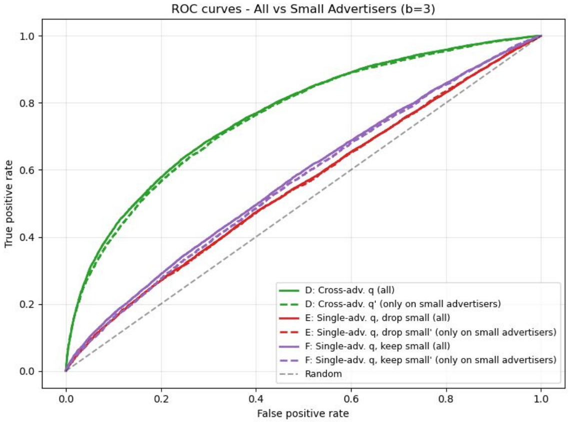












Outline

- Preliminary methodology
- *Very* preliminary results
- **Next steps**
- Your feedback

Next steps

- Preliminary experiments do not operate on Attribution: no per-device budgeting; impression features known to adtech at conversion time.
  - Move onto a more realistic evaluation with individual privacy budgets (maybe even quotas, see our updated Big Bird paper).
- Preliminary experiments investigate very simple bagging policies, params.
  - Investigate more bagging policies, such as over time, to make sure we have the best LLP model we can develop.
  - Investigate best bag sizes: want them to be small for LLP but not too small for DP. There may be a “sweet spot” that we haven’t yet found.
- Other approaches than LLP (e.g. weighted aggregate logistic regression, other suggestions?)

Outline

- Preliminary methodology
- *Very* preliminary results
- Next steps
- **Your feedback**

Your feedback on:

Evaluate optimization use cases on:
single-advertiser queries (Attribution Level 1)
vs. cross-advertiser queries (envisioned for Level 2)