Private Advertising Technology Community Group Minutes - 2023-10 Meeting

Setup

Click Here to find QUEUE! —→ Queue ←— Click Here to find QUEUE!

Note your participation here! —→ Participant List ←— Note your participation here!

ZOOM Link:

Zoom link for both sessions
https://w3c.zoom.us/j/82659868398?pwd=R2wyMlVzVGcwcmZJb1BpZmdDc2crUT09

Slack invite: https://www.w3.org/slack-w3ccommunity-invite
Our group is called “private-advertising-technology-cg”

Scribes

Willing Victims (Session 1):

- Martin Thomson
- Nick Doty
- Ben Case (willing for #148, #154, #155)
- Paul deGrandis

Willing Victims (Session 2):

- Charlie Harrison
- Michael Kleber
- Paul deGrandis
- Erik Taubeneck

Agenda

Day 1

<= 10m: Hellos, Intros, Google Doc, Call for Scribes

Policy Slides

=> 20m: Complementary results on Private conversion measurement using label-DP - #148

=> 60m: What data do we need to calculate the impact of use at scale? - #153

=> 30m: WALR - Weighted Aggregate Logistic Regression #154

<= 60m: Experiment to better understand the effect of DP on advertising decision making - #155

Day 2

<= 10m: Hellos, Intros, Google Doc, Call for Scribes

<= 20m: PATWG Status Update

== 15m: 2024 Meeting Locations

=> 10m: Update from the Empirical Privacy Metrics task force

=> 60m: Update on Steel man of proposals.

=> 60m: Review Baseline Requirements for Private Measurement

Logistics

Agenda

https://github.com/patcg/meetings/blob/main/2023/10/24-telecon/README.md

Zoom

https://w3c.zoom.us/j/82659868398?pwd=R2wyMlVzVGcwcmZJb1BpZmdDc2crUT09

IRC and Slack

Use Slack for the backchannel please. Do not use Zoom chat.

https://irc.w3.org/?channels=patcg

https://w3ccommunity.slack.com/
Our group is called “private-advertising-technology-cg”

W3C Read All About It!

Policy Slides

Session 1

Hellos, Intros, Physical space ground rules, Google Doc, Call for Scribes

<= 30m

Policy Slides

Same policies and read all about it! Smaller group and simple agenda today.

Complementary results on Private conversion measurement using label-DP

https://github.com/patcg/meetings/issues/148
Presentation: ?
Presenter: Maxime Vono

Scribe: Martin

This is a follow-up in response to a presentation from Google at our F2F in London. Followup on presentation on label DP like Google had. Already presented to google and Meta

- Summarize Google's last presentation. ML training with label privacy. X features. Y labels. Do gradient descent. Instead of using full DP on both features and labels only do it on the labels. Here is the definition of epsilon-label-DP. Neighboring datasets just differ in the label.
- One way to do label DP is RR. do a random flip on the label with DP parameter. Feed into the training as usual.

Train a model using input data consisting of features and labels. ε-Label-DP only involves noise for labels. Applies to any definition for grain. Binary labels that use randomized response at a probability based on the value of ε.

Shows improved AUC loss with higher ε. Higher caps show significantly worse AUC as that relates to having larger amounts of noise.

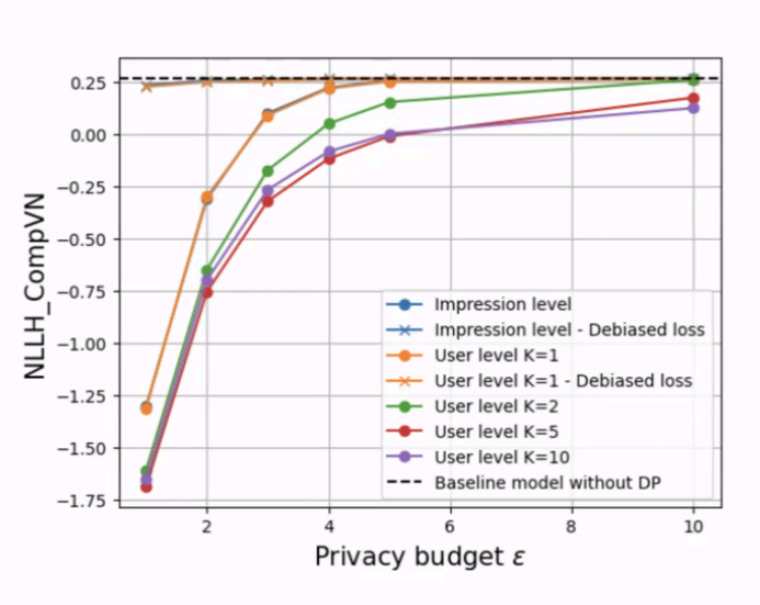
Findings

- AUC is not the best metric to use. We will use relative uplift relative to log loss.
- Without surrogates, this has poor performance for ε < 3. A number of techniques can be used to improve performance when there is higher noise.

    - Debiasing function from “Learning with Noisy Labels” Natarjan, Dhillon, Ravikumar, Tewari. Works well with log loss.

Results

- Used Criteo dataset of ~6m items.
- Logistic regression model.



- Uplift compared to extremely naive model that predicts mean of training data with no noise
- Uplift for debiasing is good. Debiased LL for contribution cap of 1 at ε = 1 is similar to higher epsilon.

Future work

- Look at other approaches to handling noise (optimal transport, etc…)
- Look at handling noisy features.
- Compare with WALR.

Charlie Harrison:

- Looking at the uplift relative to the mean of the training data. Confirming that the mean is computed with no privacy. [answer: yes, just ignore that and look at the absolute log loss]
- Without surrogates the loss is poor when ε drops below 3. So at around ε = 3, the performance tanks, is that correct? [answer: yes]

Nick Doty :

- You are calculating the current performance and comparing these different models. What are the conditions for performance in the dataset? When you calculate the baseline are you assuming a comprehensive and correct view? [answer: without considering 3p cookies, I took an initial data set of features and labels that was clean (this is anonymized, but public); the baseline model is trained on the clean dataset]

- I'm not sure that this is the best baseline for comparison in this context. [answer: the goal is to compare performance with and without DP] We might want to compare against other baselines in terms of understanding the commercial impact against likely technical privacy protection outcomes.

Ben Savage:

- If we ignore all the debiased ones, what about the two best performing lines. I don't have a lot of intuition about what the difference between the best performing is.
- At ε = 1, going from .28 to .23, how bad is that? What does that matter for a business. [answer: if you have a 1% improvement, that is huge for business metrics. 1-3% can translate into huge improvements in online metrics; this is an offline metric. A jump from ε=1 to ε=2 is massive here. Even the tiny delta between ε=4 and ε=5 is significant.
- Which would be close enough to what is in production such that it might represent a tolerable loss? [answer: considering the online metric, the drop is still significant]
- [note: this is only offline performance]

Aram Zucker-Scharff:

- Are you looking to match labels from 3p cookies, what are you measuring in the dataset? [answer: (notes summary: The capacity of the model to predict the user's likelihood to convert). performance is based on conversions (sale/not) and the goal is to predict the probability of a sale, which is ultimately compared with the actual event. So if you predict p=0.2, and the truth is either 0 or 1, log loss is a metric you can compute.]

Charlie Harrison:

- Can we look at the zoomed in version of the top 2 items, perhaps framed in terms of percentage loss? Hard to squint at these.
- Responding to Nick about what we should use as a baseline. My philosophy is that we should measure performance in this way, but we should not necessarily decide to move forward or not based on how it performs relative to having 3p cookies, but when we are comparing approaches, this true baseline (with no noise) is the right baseline to look at because we want these things to be as accurate as possible before privacy mechanisms are applied. Ideally, we get the most optimal thing compared to ideal. Otherwise it will be hard to compare mechanisms without having a no-noise baseline with which to compare.
- Want to know what Nick was pushing at. More juice for the same privacy, or something else?

Nick Doty:

- I don't have a problem with this baseline. It's useful to describe things this way in comparing algorithms. But it might be useful to have other baselines with which to compare the business impact. Measuring the business impact relative to perfection (or to no privacy protections ever coming into play) would be delusion.

Charlie:

- OK, maybe there are other places to consider the business impact.

Nick:

- Or maybe what does this look like when considering a system that is compliant with EU law or other similar sorts of constraints.

Charlie:

- We should get as close to ideal as we can privately, but we should not use that ideal when making a final comparison because one will be non-private and another won't.

Ben:

- From my perspective, there is a training data set of some historical data. Ads were shown to people. Some converted, some didn't. We had good measurements, because we had 3p cookies. We're just comparing the performance of the model against the collected data.
- Are we evaluating the model against the data it was trained on? [answer: No, a train/test split has been performed. All algorithms (incl. The baseline) have been trained on the training dataset and tested on the testing dataset.]

Aram:

- Good to measure this, but this is not the only measure we should be considering.
- Considering the ability of the models to generate metrics and not necessarily the model's ability to measure business success. Not that this isn't great work, it is, but it would be great to look at business/revenue impact more directly.

Michael:

- When you talked about the model performance degradation in this offline analysis and how that translates to business outcomes…there are two ways in which the model could degrade business outcomes. Which is it?
    1. ROI going negative. That is, advertisers buy when the ad doesn't return that investment (in the aggregate).
    2. You are worse at predicting which auctions are the most profitable to bid in. Therefore, you lose out in auctions to others who are better at predicting outcomes than you.
- The first is valuable, even if no one else were present and bidding.
- The second might not be as relevant if everyone else suffers a similar loss of predictive capacity.
- Can you explain why this accuracy going down causes bad business outcomes. [answer: not an expert in bidding models, but I can investigate for you]
- The big question then is, in translating from log loss to real performance, do you have a different guess about performance in auctions where all other bidders are operating under similar constraints.

Thomas Prieur:

- Our clients measure revenue in terms of ROI. Eventually reductions in advertiser ROI translates into reduced publisher revenue.

Alex Whitworth:

- Similar questions. How representative is logistic regression to production predictive systems? [answer: logistic regression is a common baseline in the industry. It scales well. ANN do not scale well. LR is a strong baseline.]
- Are there online advertisers using LR in production [answer: yes, completely]

Joey Knightbrook:

- Can you share slides? [yes]
- What was the conversion rate in the dataset. [6.74%] Wow, that's high.
- Need to ensure false positives don't overwhelm true positives. I wonder how a rate closer to 1% would affect model quality. [answer: we could look at other suites with lower conversion rates, but there are tricks to adjust datasets so that effective conversion rates are higher for training purposes]

Charlie:

- Want to put this in context. What would an API look like that generates labels like this.
- This goes back to the presentation from a few meetings ago, where I described a per-event based set of data, with noisy labels (or randomized response). That is the sort of setting where this would apply.
- In terms of the APIs under consideration, almost all have the capability to do this. IPA has made some modifications that might make it possible to make queries on individual impressions. ARA has event-level reports. We don't even aggregate at that point, local DP can be used. PAM publisher reports includes something similar to event-level reports. You could - in theory - make individual queries across every impression, to get a noised label for each impression. PAM doesn't include randomized response, so it would use the Laplace mechanism, but you could turn that into a binary classifier with post-processing. I have an issue open on PAM, in case this was not intentional: https://github.com/patcg-individual-drafts/private-ad-measurement/issues/4 (Luke took an action).

What data do we need to calculate the impact of use at scale?

#153

Charlie defers until next time, still examining all the possible requests.

High level goals of this workstream. We want to know what the absolute scale impacts are for different API decisions that we make. In the comparison table, each showed different scaling properties. Computational cost and bandwidth are all over the place. To compare, we'll need to have a representative dataset that includes data for real applications. Then we can determine the cost of each of the proposals.

That means being thoughtful about the data that we ask for. e.g., in PAM. There is a tradeoff in the Prio backend, which means that the number of buckets in the histogram determines the size of submissions, linearly. The question of how many aggregation keys you need is therefore necessary to answer, but difficult to predict. There is how many you use today (in 3p cookie world) or in ARA (in chrome). But the answer might differ for

different browsers. For low epsilon, you could maybe have more buckets, but high epsilon might need more values in each bucket. A lot of these parameters might change in use.

We might use a non-private baseline, but that might not be representative. Let's look at distributions for different actors.

A separate track is what cost is considered "reasonable". This asks Michael's questions. Is ROI negative? Trying to get a sense of where costs could be is interesting. Some of the IPA presentations looked at the cost as a fraction of campaign budgets, but other approaches might be useful. Some of this data might be sensitive, so we need to think about it a little more before putting something out.

Nick:

- (If this off-topic, let me know.) One assumption is that we would have different parameters for each browser vendor. And campaigns would be segregated by browser vendor and run separately. It seems like that will increase noise and decrease the value for visitors from smaller browser vendors

Charlie:

- when we talked about this before, a layered APi would allow different browser vendors to make different choices. Yes, it is likely that different vendors might support different levels. There might still be value in an interoperable backend. Maybe you could spend a common part of budget (intersection) to answer common questions, then use the spare for the more generous browser vendors to answer other questions.
- We probably can't agree on a single value across all vendors. But we might have interoperability that would allow for cross-vendor queries.

Ben Case:

- agree with performance aspects of number of histogram bins. one additional dimension: when we have an attribution, you may want it to contribute to multiple different histograms, even when each doesn't have many bins. capture the input for API users on how much they want that.

Brian:

- thanks for taking this up. important to be precise in describing 'performance'.

Charlie:

- a table with the measurement and an explanation of what it means, like what the costs are for bandwidth and computation for each. and how we can compare API proposals on those costs/performance.

Martin: Queued with - to Charlie's point, I consider it a failure if we don't have interoperability at least to the level of common epsilon budgets. More generally, this sort of exercise is ultimately not going to be possible in this setting (i.e., standards). We'll do what we can, but at some point we need to start breaking some eggs.

Interoperability on the backend is important/essential. you should be able to query across all the browsers with that budget, even if some browser has additional budget. you might also want to do cross-tabs by browser, but that's separate.

Useful to have some understanding of the scale and impact. But at some point, have to start running code and evaluating what the impacts are. We'll never know the exact commercial impacts on every party. Should time box, lest we talk about this forever.

Charlie:

- Yes, that sounds fine to me. Just want to see a little bit more.
- Primarily interested in the question of economic viability. We can improve MPC performance by adding computation and communication, but at some point the cost is so high that we should just do the simpler, less performant thing. I don't have a good intuition about the comparisons yet, or the impact of a more maximal feature set.
- Could even suggest a set of metrics that organizations can run themselves in private, and then share a stack-ranking or overall impressions.

Martin: an approximate stake in the ground would be useful. Some analysis will be done in private with data that companies won't be willing to share.

Mariana: ask everyone what's on the wishlist for parameters and features vs. encouraging people to think about how they would use a limited number of queries or buckets. Does Safari/Webkit have a sense of what they think is a reasonable number of buckets, and then ask others whether that seems feasible to advertisers.

Aram: true economic viability: we can't ever measure it fully in advance. how the marketplace will adjust in terms of spending and what they think is worthwhile is difficult to predict. Google and Meta have particular insights into their marketplaces, but different organizations will have different requirements and priorities. hard to predict what buyers will be willing to spend on, even if we asked them now. time box, because could get different answers forever.

Thomas: we cannot ask individuals and get a clear answer. prefer building a test facility and a test dataset, and let companies run it themselves and get some test results themselves. ideally get it so that companies will be willing to share their test results. if we just ask, there will be very different opinions from browser vendors, users, adtech. Prefer a test.

Charlie: not sure this requires something so in depth. just order of magnitude of what impressions you serve, etc. maybe these are all tied together, but can survey some just for scale analysis.

Aram: trying to compare more private and less private Internets is not going to give us good results. because buyers are going to continue buying the less private access until the switch is flipped. don't expect advertisers to just suddenly stop spending money on advertising, but expect shifts that are hard to predict.

Thomas: we have some data on that with some browsers that have made privacy shifts. money will go to different places, but we don't think it'll go to publishers or at least small publishers.
Aram: spend moving away from private browsers doesn't predict what will happen when all browsers are more private.
Thomas: clients don't care about the channel, just go wherever the ROI is best.
Aram: I understand that but I'm not sure they really will understand what the shift in ROI is until it happens.

Brian: adtech is always dealing with complex data that is hard to predict. but we can get a sense in the room of is-this-workable-or-not. +1 to Charlie's plan, and getting the dimensions we should consider, but don't think we can/should get so precise.

Ben: take a use case that needs to persist, and then develop a minimum viable product that will provide that use case. (predicting budgets and revenue, etc. is too hard.) how many websites do you run campaigns on? we need to know that. each website/breakdown increases the number of multiplications. and what data needs to be loaded on the device, do you need to download the ad campaign x website breakdown?

Charlie: more of a smell check, or order of magnitude check.

Graham: distinguishing between required vs nice-to-have. looking at ad reporting interfaces might be a useful place to start. what are the default dimensions that platforms provide to customers, as distinct from what is available in advanced settings.

Alex Whitworth: number of breakouts is of very high cardinality. any users can contribute to dozens or hundreds of statistics for any individual campaign. we have a proof of concept, planning to build a production system. but the hurdle is moving from hundreds of statistics per user, needing to limit that to a much smaller number in order to provide a reasonable privacy guarantee.

WALR - Weighted Aggregate Logistic Regression

#154

Scribe: Ben Case
Slides:
https://github.com/patcg/meetings/blob/main/2023/10/24-telecon/Weighted%20Aggregate%20Logistic%20Regression%20(WALR).pdf
Presenter: Ben Savage

Writeup: https://github.com/patcg-individual-drafts/ipa/blob/main/logistic_regression.md

BenS: present a cool result from some researchers at Meta came up with. We put it on the github, see issue for link.
Going back to discussion, what we are trying to do in optimization/calibration. Want a function that predicts the likelihood of conversion given an impression. Assume some data features about the ad being shown or person shown to. Assume the party showing the ad knows x. Different than fensed frames, FLEDGE. This is the case that info is known.

Logistic regression – WALR is a way of training. Same approach Maxime presented on earlier. Not a complex ML approach. f(features) is sigmoide(features dotproduct model parameter), here features = x. Just adding up these weighted features to get a score. Then use sigmoid to get a value in 0,1.
Simplifying assumption we can use binary features. Maybe need to convert a feature into multiple binary features. Will make easier to analyze and implement in MPC. often doesn't really change things.
What is LR? Once trained with get this list of numbers as long as list of features. Tell how correlated feature is with the likelihood of a conversion. Large greater correlation.
Derivation of WALR. some math to come. How to train LR without concern for privacy? Use gradient descent, define a loss function, initallize theta to something random. Compute gradient which tells you which way to change to be less bad. Find direction "down hill" update model params by taking a step in that direction. Just keep doing this till can't do better.
Here is a loss function – average loss across cross entropy loss. If prediction is closs to correct, small loss. Gradient of this loss function. Proof to the reader. See slides. Observation is there are two terms. The first term includes theta (model coefficients) and x (feature vector) – no part of this is private data – it can be computed in the clear. Don't need MPC or anything. The second term is average over labels (private) times the features – this is the only part affected by the labels which is the only thing in this setting that is private. There is no theta in this term. So each step of gradient descent this never changes! We can compute this term once and do the rest of it outside of MPC. compute this value once in MPC, add noise to make private. We end up with a noise gradient – which is still mostly in the right direction down hill. We'll call this second term the "noisy-dot-product". Math done.
WALR with this – compute noisy dot-product. Continue as before staring from a random theta and going in the right direction with a step.
Possible to train a LR model with in this label privacy setting in a very optimal way. Very little needed in MPC. the rest of the work outside of the MPC.
Means we only need to ad noise one time. Compared to label-DP, that is local DP. but this is central DP, so the more rows you have the better the signal to noise ratio.
What exactly computed in MPC? Feature columns and one label column. In MPC some labels are 0 and some 1. What you are doing is adding down columns where only adding in ones with label 1. That is all, just a bunch of conditional sums, really counts. What you get is a total per feature. Then add noise.
How much noise to add? The maximum change in L2 norm of that sum. L2 norm is to take that k dimensional vector – square each and sum then take sqrt. If one label flips the change at most is one row of all 0s to 1s. At most this is then sqrt(k) on the l2 norm. So this is the sensitivity for the DP noise. Has nothing to do with number of training examples, so signal to noise ratio can get really good.
Compare WALR with what we've been discussing of noisy label RR approach. WALR lets us do computation where output is an aggregate. We all haven't been able to come up with a mathematical reason we like aggregation vs not, but many folks likely do like aggregation more.
For calibration, this might be enough.
Intuitively, we don't have data yet – but Criteo offered to investigate – but we'd like to see that for a given epsilon we get better utility.
Also is is more purpose limited than noisy labels – you can do anything later with that data – pro/con of it. Maybe we don't want so much reusuablity of what you can do with the data that comes out. Ready for questions

Charlie Harrison (same approach as the winning solution in https://arxiv.org/pdf/2201.13123.pdf? noisy dot product vs. noisy histogram):

Seems similar to winning solution to Criteo competition. Is this the same? Also winning solution by FB

Ben: this is a different approach but also by Meta. we think WALR is better

Charlie: same setting of features private, would help to have comparison with that paper.

Charlie: other question, might need to see slides. The MPC output is noisy dot product or noisy histogram? Is the noise similar to these two outputs? Wonders why we need to change IPA at all? Why can't we compute with existing mechanism where we get out a noisy histogram. Noisy individual feature sums and then multiple outside with the featur

Ben: you're saying could do separate queries for each feature column. Probably just an efficiency difference.

Charlie: we've talk about impression with multiple breakdowns keys. Seems similar but maybe less efficient

Ben: yeah the same i think but probably less efficient. Here also assuming binary features which is easier than general IPA query where not binary. Implemented already inMPC

Charlie: always assumed IPA could do this paper approach. So wasn't sure why needed more code to do it. But value in letting one impression contribute to many outputs but can see there could be specific optimization for this. Aside curious to see Criteo competition with this system, 200 features, see if can handle the scale.

Ben: should be able to handle would be 200 multiplies per row.

Charlie: were not all binary features so would need to reencode. That would be interesting

Ben: good point

Mariana: ask if pointer for writeup, was confused if the same budget as label DP approach. Presentation seemd would be better but discussion with charlie was confused. Are we getting better params than just releasing labels?

Charlie: label-DP will scale with number of rows, but with this only adding noise once so in principle should be better. For LR this would likely be better. Advantage of label DP is you don't have to embed training into the private compute framework – maybe other DNNs. that is benefit of label DP.

Mariana: my question was about budget consumed. Also writeup?

Ben: we have writeup.

Mariana: think that it is because loss is linear in derivative. Should apply to other models

Ben: would love to see, have been looking into but not seeing it come out

Ben: we could do this on-device with PAM or ARA. nothing IPA specific in requirements

Nick: share intuition aggregation has privacy advantage. Central DP has strong utility advantage. Genuine question, what is implication of features if they contain unique ids for users, could you learn about the user? Is there a sensitivity to the features we have to think about?

Ben: most extreme thing you could do. Can't easily encode unique id into binary. Could try a one-hot encoding with 1 for that feature if Martin otherwise 0. So could be done that the aggregation in the end is only capturing one person's data. Probably not cost effective since would need one per user to do at scale. More features you have the more noise you add. But the final backstop is the DP. we have to pick epsilon that makes us happy even if a feature singles out one user. Don't think people would actually single users out like that and get any good performance. Could have a discussion about what the features are and should there be transparency around what they are. But if you look at cost in mPC as linear in number of features so costs more in MPC and more noise, so natural backpressure about having lots of features.

Nick. helpful

Charlie: concern features too sensitive is valid concern. In GPS we have proposed mutations but as platform only so much we can do. Site that knows a lot about you like FB it is hard to know what the features are they are serving ads based on. What we can do as platform is make sure we limit features that are based on cross-site information. In topics we limit features based on cross-site features. In that setting of protected audience these approaches don't work since not all features know in the clear. A more advanced setting for model training. We could present more on this setting; but those APIs aren't usually discussed in this group. Splitting features into sensitive vs not.

Ben: in world today with cross-site tracking, features are likely derived from cross-site tracking – knowing the person previously purchased something on that site before. But in a world without cross-site track you wouldn't have that feature. So back to our previous discussion where when teh future changes everything changes. Wouldn't be able to an apples-to-apples comparison with the same features since you might loose some of the features.

Nick: yeah, that's to my question. Is it that stopping cross-site tracking removes these cross-site features or do we still need to consider in the protocol. What to know the privacy protection we can provide if features are sensitive and could have come from non-platform ways like PII.

Ben: previous answer than not scalable to many features. If there is some out of band want to get cross-site feature, then could still use it and likely would be predictive. The other category of sensitive is if based on some protective feature. E.g. meta settlement not to use certain features in ranking ads for housing and employment. Legal compliance. Making sure these features not used indirectly ect.

Joey: similar to mariana's question – have you learned about extending this to more complex models? Can't you create features for cross-features? More on this conversation for complex models?

Ben: as mariana pointed out the gradient of loss function of LR is nice. Only latter term affected by labels. My understanding is that is possible because property of LR model. When you take derivative is nice with chain rule. If not linear then will contain theta in the second term. My understanding of worked her but not more sophisticated. I'm trying to work out for two layer NN. generic extension is DP-SGD – won't be as simple in

MPC. but would let you train more complex models. Happy to investigate that if interest in more complex models. Could maybe pick a different loss function. Give as problem to the mathematicians.

Joey: think you might lose something with non-independence. But could use features to create features which are products of features.

Ben: yes you can push LR pretty far. Could add dependent features. Want to share and let folks do with it. Test it and see performance. Maybe best approach is to get something simple to start.

Aram: good note to end on. Moving to Erik's presentation but questions next time

Experiment to better understand the effect of DP on advertising decision making

#155
Slides: The effect of DP on (ads) decision making - PATCG 20231024 / GitHub version

Erik: hope to get feedback so good to let folks think about it. Will add public copy of slides after.
The effects of DP on (ads) decision making.
Good discussion earlier –this is a subset of the goals.

1. Prevent user level tracking
2. Enable agg measurement for cross-site behavior
3. Design something useful in meeting both of these.

Privacy vs utility graph – frontier curve. Difficult to understand where to stop can always slide around the curve. How to come to a tradeoff?
Want to establish some bounds – lower and upper bounds when utility is meaningless and on privacy when privacy is meaningless.
Goal for experiment is to understand the lower bound on utility. Some lower bound on epsilon. Epsilon > x or not good enough utility.
Simple example – binary decision to make. Make input privatized. As epsilon ->0 decision should be come random. If we can figure out what this curve looks like
Experiment design

- Cause effect of adding DP to outputs of an ads campaign.
- Creates 10 outputs, then copies and ads noise
- Shuffles all 20
- Lets you decide to increase or decrease spend.
- Checks how many times you made the same decision on noised vs non-noised pair.

Configuration params

- Campaign size
- Conversion rate

- Variance o

Simulate conversion counts from beta-binomal process, so each campaign has counts draw from binomial with its parameter drawn from the beta-binomail
Let you see if you want to increase/decrease variance. Started variance off of criteo dataset.
Too simple? Usually the decision you make is based on some simple rule. You could just simulate based on a decision rule. You see the impact of DP on that decision rule. Can make correct decisions with larger epsilon.
Are we learning anything interesting from playing the game? Probably no.
If we choose different decision rules, we can see some interesting things about what people's decision boundaries are. Would help to know what people are looking for when making decisions.
How can we make this better? Alex had to leave but opened issue to maintain or not enough data. With noised value could have confidence interval on the true value. Could ask question about what to do with this result.
Making different bad decisions can also be simulated, maintain vs increase, vs decrease.
Different decision rules can see some gaps here. In the face of uncertainty what do you do? Can you still make decisions?
Other ideas – add more user configurations. Amount of noise here is large, epsilon is small but that is assuming one single measure on a large group. Not the usually case. If you add up to how much you'd want to do, gets similar to the question we were asking earlier. Maybe can be incorporated here.
Can think about splitting spend and making daily adjustments, like lemonade flash game
Did you make the same decision or not to how many conversion did you get or not. When does your decision making process fall off a cliff.
Looking for folks to actually run this and give to recruit people. Want to make this compelling to this group and get your input.
Extras: simulation on individual events. If you use same median decision making rule can see probability you get things wrong. Epsilon 3, 11% wrong. Looks like earlier curve. If we can come up with larger groups of people, say 1000, then will be to the right of this curve. So maybe there is is space on the frontier between our lower and upper bounds.
Ideas welcome!

Aram: if folks have questions put in the issue linked. If you presented make sure to upload on github. Sign in

Charlie: last point about simulating individual events?
Erik: if you had a privatized value and true underlying, if you have a rule then you can look at how often you are wrong.
Charlie: make sense. Like simulating RR with Laplace mech.
Erik: yeah, similar to differencing attack on aggregations with Laplace
Charlie: main question is if we decide that private decision making is right way to consider this problem, you can always design a more efficient privacy mechanism that is exactly the decision making rule. Could design a mechanism that is stack rank these campaigns. Probably the case in some use cases.
Erik: two things – think over simplification of what the decisions are. This lower bound on what we'd want to do.
Charlie: if we want output of privacy decisions instead of dashboards then there is a different way to do it.
Erik: hopefully we can get to a place where we can spend budget on different things for your use case
Erik: other thing was trying to do here is to make this like an arbitrary decision instead of a decision rule. Once you have a rule you can just simulate and survey is not interesting. A survey may reveal interesting decision rules.

Aram: upload slides. Questions for Erik on the issue please. See you Thursday same time. Same place .

Session 2

Day 2 Agenda:

<= 10m: Hellos, Intros, Google Doc, Call for Scribes

<= 20m: PATWG Status Update

== 15m: 2024 Meeting Locations

=> 10m: Update from the Empirical Privacy Metrics task force

=> 60m: Update on Steel man of proposals.

=> 60m: Review Baseline Requirements for Private Measurement

Hellos, Intros, Physical space ground rules, Google Doc, Call for Scribes

<= 30m

Policy Slides

Same policies and read all about it! Smaller group and simple agenda today.

PATWG Status Update

Scribe: Charlie Harrison

- Sam: Charter is out for AC review , sometime in November. Encourage your AC rep to vote on that
- Aram: We have 32 participants, would like to see a good turnout for voting. Talk to your AC rep to vote. I would love if you vote “Approve”
- Sean: Just want votes. Silence is worse. Any heartbeat is helpful.
- Aram: More votes > less votes
- Sam: We are looking for at least 20 “yes” votes
- Aram: Any qs about the chartering process?
- Brian: the proposed PATWG charter on the web?
- Aram: They will have also received a link to this in the AC email newsletter. If they don’t it is under the PATCG github org.
    - https://github.com/patcg/patwg-charter

    - There is a Call for Review of a proposed charter for the Private Advertising Technology Working Group: https://www.w3.org/2023/10/PROPOSED-PATWG-charter.html
    - Please have your AC rep review the charter and indicate your support using this online form: https://www.w3.org/2002/09/wbs/33280/PAT-WG-Charter/
- Miguel Morales: Plans to make this a working group? Relation to Privacy Sandbox? Merging the sandbox and the work being done here
- Aram: W3C charter is our process of spinning off a working group from a community group. It is common for there to be a community group that is public facing that does the work of initial incubation of proposals. The working group (PATWG) is the body within w3c that can create standards. We are going to be running both groups simultaneously. The WG is for members only. It's good to give both. In terms of Privacy Sandbox, the way we have chartered it is that it will take up the work of measurement. This is the first thing that will likely go to the working group. It could pick up additional proposals. We have a few of the privacy sandbox proposals from Google in the general incubation space. Not really the work of the group but within our "us keeping an eye on it". If it is up to proposers to bring the group their proposals. While we would assume that proposals that have to do with Private Advertising to come to this group, that is up to the proposer to move it.
    - I have no specific knowledge of any other proposals coming to us in the future, but hope those proposers will come
- Miguel: that answers the q
-

2024 Meeting Locations

- Sean: proposing the same schedule as last year
    - Not finalized or looked for conflicts
    - We know that TPAC in California in Sept
    - We want a F2F in Feb and June. We have strong offers from Singapore, Boston, Cologne. Happy to go anywhere but want to "spread the pain equally" on travel budgets
    - We have not yet been to Asia
    - Throw it out there for discussion
- Erik: Offer Seattle. We have a meeting room outside of security. ANother west coast option
- Sean: heard from people they don't want to go back to California.
- Rachel Yager: Offer NYC. New to this group and an evangelist for NYC. Location Midtown, east/west some convention center. Happy to offer support for Singapore as well.
- Aram: our upcoming TPAC is in Anaheim California. Probably want to avoid having a F2F near there (geographically)
- Rachel: the following year I saw mention of Singapore for TPAC.
- Sean: Great, even more options. Maybe just one F2F between tpacs?
- Aram: just meant not doing *california*
- Joey: July almost no one showed up. Does it make sense to have a July meeting?
- Sean: Happy to drop that if no one will show up. Then we'll have back to back F2F
- Brian: Given the probably unusual nature of next year, propose we keep it.

- Sean: are you referring to the WG?
- Brian: yeah and cookies going away
- Sean: when we have the WG, we will shift this time frame to that.
- Paul: Support for Boston as the option. Happy to do legwork to help. Boston in Feb is the most unpredictable month for weather.
- Charlie: When will the WG meetings start?
- Sam: Too hard to answer
- Kyle: Another +1 to Boston. Really nice conference room locations. Feb weather is bad.
- Aram: In the chat, June is the opposite situation - hotels filled w/ college graduations
- Erik: Do a poll?
- Sean: Yeah. Everyone needs to consider their budget, but asking people to be flexible
- Aram: Setting up an issue for this.
- Rachel: Feb is lunar new year, so Singapore might not be available.
- Ben: Let's do after Chinese new year
- Rachel: Yes, Singapore after Chinese new year Feb 10, so after Feb 20 is fine.
- Rachel: New York City is good in June and July, and also in Feb (even if it's winter).

Update from the Empirical Privacy Metrics task force

#137, doc

- Tamara: Our plan was just to have one last call for feedback on this proposal. We haven't made a ton of progress since the London meeting. We do have a document.
    - Empirical privacy comes across as all encompassing. Changing to something like "comparable" privacy metrics
    - Taking suggestions
    - Love feedback on the goal of the work. In the document itself, so we can keep this short.
    - Lots of other constraints in the doc to make the problem more solvable. Lots of good conversation here about what makes a good metric. Concrete feedback would be "of these, is anything missing? How would you stack rank them" [in ref to "Properties of a good EP metric"].
    - Lastly, we have a project plan. One last ask for volunteers. We are planning on meeting a week from Friday to finalize. We already have one volunteer creating a dataset. Two milestones, and present metric + research results to working group and solicit feedback.
- Nick Doty (membership inference vs re-identification. or what should the top threat be for evaluation?): Thanks for the overview + organizing. If you are thinking about this area, what is the top most privacy threat you are interesting in evaluating in a metric. Proposal here would be "membership inference". Also heard "reidentification". Some other threat?
- Tamara: Started w/ membership inference, but Ben pointed out we could do reid in this scope.
- Ben Case: There's two stages. One is aligning on the definition of the metric. That's easier than evaluating each proposal. Hope in the nearer term we can write down a set of proposals. Want progress so scope it down to membership inference first. We can do that for more metrics
- Tamara: more shameless plug for volunteers

- Martin T (do we need to ask about expected privacy vs. worst case?): Hope the proponents of the various proposals can help you. How do we think about this from the POV of privacy at scale vs. worst case. Some discussion earlier this week was talking about “worst case” but contextualizing in “for how many ppl is it the worst case”. Membership inference seems like a difficult target to attain. # of people who potentially fit into the worst case differs.
- Tamara: For every impression, you could have some probability of identification / inference. Still thinking through
- Ben Case: mental model for metrics: how will we measure “budget” in these proposals. These metrics will want to compare across budget metrics (e.g. DP). Average case, etc. to get something comparable.
- Martin: we will need to do discussion around the nuances
- Ben Case (good to get folks input on the use of auxiliary data): Want to keep the scope of auxiliary data to be manageable. We want to mainly look at what the impact of the API on the web. Not that there exist other ways of tracking. Don’t want to scope out all other relevant forms of data. If you have strong thoughts on what aux data should be considered in scope, open to feedback.
- Mariana: Membership inference. Are we intending to look at what happens after we train ML models? This seems quite intangible to me. All possible ML models people might be training, probability of membership inference on those. Second question / thought: are we going to try to give an interpretation of these metrics? If we go w/ “budget” epsilon, are we going to try to understand what different values mean in practice.
- Tamara: don’t want to re-invent DP. Our goal is the metric can be described in one sentence. That’s one of our targets. We’ll keep this scope to attribution. Still worried this might not be productive, we need to see if we see value in it.
- Charlie: I guess I need to read the doc I’m a little bit behind. One thing I still don’t have a firm grasp on is what the threat model is for this metric. Especially if this is empirical and we’re trying to generate datasets is the goal to generate data sets and understand conversions or are we trying to get a dataset of user activity on the web and unleash an adversarial actor and learn about the user’s web activity to understand the impact. One is more biz as usual you are logging legit conversions and the other is you are really trying to abuse the APIs where you are using them for surveillance. Which end of the spectrum you are looking at will really dictate the interpretation of the metric and also the question of how we will analyze it. Ad engagements may not be the right input.
- Tamara: Yes I was thinking more of the first case. Trying to understand values of the epsilon. ARA vs some other privacy unit. Can we explore the utility of that data. I understand it won’t be comprehensive.
- Charlie: :This won’t be a worst case metric but also this won’t be a worst case adversary, no click jacking no ideas on how to hijack users is to be incorporated here.
- Tamara: Yeah, but if you have ideas we are open to looking at that in the future.
- Charlie: I don’t have a strong idea on how to do that here, but I just want to be clear that we are on the same space.
- Brian May: It’s been a while since I read through this doc. Think about privacy as having two sides. Trying to learn something about a person, and using information you have to impact their decisions / impact them. I’m assuming this is mostly on the first aspect
- Tamara: one of my worries. Most of the APIs try to allow population-level insight. THe line is fuzzy because population-level insights looking like a membership inference attack
- Brian: what is the potential that the learnings from the API help me influence user behavior
- Tamara: if it allows specific user insights, not good

- Brian: Is this scoped to a specific time window, or all time?
- Tamara: we haven't decided. Ideally this will be a composable metric
- Charlie: https://github.com/patcg/docs-and-reports/blob/main/design-dimensions/Dimensions-with-General-Agreement.md#inclusion-of-a-time-or-interaction-dimension
-

Experiment to better understand the effect of DP on advertising decision making (cont'd)

scribe: kleber

#155

1. Unit of measurement
2. Establishing a lower bound
3. Ideas for extensions
4. Help recruiting participants for a study

Erik Taubeneck: Four things that I'm looking to get feedback on. See issues above. First. how this works right now: The unit of measurement that then gets noised. Epsilon is applied to each measurement, so getting an estimate of how much utility is impacted by noising individual measurements. These proposals actually all have a budget which you are dividing over all of the individual queries, which has more complexity. A simple tool to give you a sense of what it would feel like might not need to bite off that complexity. Different proposals construct the budgeting in different ways, so it's not necessarily easy to modify this to take into account the different ways proposals all work.

Charlie: Campaign-level privacy unit probably not very realistic. Adding more complexity here might not help with the goal, answer the question "how much does noise alone impact utility?". Opposite direction maybe: Don't think about epsilon, just vary std deviation of distribution of noise? Then you can back-figure something about how many campaigns you run, calculate epsilon as a derived output in lots of different settings.

Erik: But with Laplacian noise, epsilon to std dev is 1:1?

Charlie: composition doesn't need to be built into the tool, could do it after the fact, and if you know "Gaussian with std dev 10" is the bar for what we can tolerate, then we can design systems around that, built your query such that noise on each campaign meets that bar, but epsilon per query can vary based on the shape of how you make and measure campaigns. I don't know that it's useful to have your tool say "okay if I'm running 100 campaigns and…" because composition models are different.

Brian May: When we look differences in how privacy budgets are managed, differences don't really matter much to people who don't spend a lot of time understanding different mechanisms. Absent any kind of common standard, maybe just looking at level of noise is all that's valuable.

Ben Savage: I was talking to someone at Facebook in Ads Reporting space, who said that a decision you might want to take is "when looking at two different creatives, you want to know which one is best". Maybe that's an interesting thing to experiment on — ability to keep ranking a list correctly?

Michael Kleber: For the list ranking thing let’s be careful b/c when we add noise you are more likely to rank similarly the elements on the list when they are close to each other. We might get the same answer for the 1st and 2nd creative if their performance is similar and that’s a good thing.

Ben: This is a var that would have to be considered in the thing. Some kind of confidence interval. How confident are you that the performance is what it is . If there is a “P” test how often do you mess up identifying the difference.

Erik: cool. Gets maybe to second issue, Charlie's comments on Tuesday. If the goal is to test some "median decision rule" or make other decisions, there are better ways that we would use noised information to make those decisions. That means we're not actually establishing a lower epsilon bound, because there are more efficient ways to make decision. The problem is that there are many different decisions based based on using these dashboards, we thought we're just picking one of them because it's a good starting place. So does the "you could have done better if you did your DP smarter" argument cut against any conclusions? should we instead try to target DP at specific decisions?

Charlie: My thoughts are nuanced. Lower bound is useful, even if we're not building it into the APIs, because it tells us "it is theoretically possible to make this decision in a good way". Depending on the privacy parameters that some browsers end up making in the APIs, I'm skeptical that it will be reasonable to make these kind of general dashboards that you can use to make lots of decisions, especially for smaller amounts of data e.g. small advertisers. I feel like we may be in the situation where building something generic may make it not useful for lots of advertisers, and so the custom, tailored algorithms may be the *only* way to get reasonable utility. "Just give me the best ranking campaign", if you're a low-volume advertiser on Firefox, maybe that's fine and that's all you can get.

Brian: I agree with Charlie — need to look at this in terms of advertiser tiers, what information they will be able to get based on sizes of campaigns they are going to be able to run.

Erik: We do include campaign size and conversion rate, so we do capture that a bit

Charlie: Note that almost all campaigns start as a cold campaign — and the "cold start" problem is really hard in DP, start with just a little budget in each of N campaigns and with low volumes and get just enough information to figure out which one to ramp up and put all of your budget in. Advertisers don't want to just put tons of money into things to find out which does well. Advertisers of all sizes have this problem

Eric: This gets to third issue: extensions. Having this to over time, get day-by-day campaign results and you make decisions. Makes this tool less simple, more like a simulation of running a campaign. You could imagine changing the tool to "day 1, day 2, day 3…" and letting you make changes daily, but once you take into account things like campaigns learning, you can get far into the complexity of real-world campaigns, and too complex for this tool. So how far should we go, what should we try to measure in the tool, at what point can we extrapolate?

Wendell Baker: Thanks for putting this together. Realism is of great interest to us at Yahoo, we're trying to figure out how this is going to work — errors and noise are scary to a business and to an employee making decisions and being responsible for them. Value of the pitch: you would *always* get to a reasonable answer, or else you would *never*. In the next few months at Yahoo we will have our systems operational and have pilot money running on these systems and we will be injecting noise and seeing what the outcome is. Something that could give a bit of training, 1-dim or 2-dim realism and pick just a few effects, would be really helpful here.

Erik: Thanks, will reach out offline for what dimensions are important.

Charlie: In the interest of starting with a simple MVP, getting some simple results make sense, but there is value in a more sophisticated setup. I would encourage you go do an experiment where you loop in real traders working for an actual agency, give them a system that gives them a view with noise. Maybe even get buy-in from real trade people who attend this meeting. Could learn quite a lot, but this might turn into almost an actual product launch — but if you learn what queries lead to what decisions, hugely valuable. That's a high endpoint in terms of complexity.

Erik: Yes, we do want to get there, that's the next point!

Luke: I admit ignorance of how people run ad campaigns and make decisions about them. Ben told me that FB had come up with a system where they won't give bad information to advertisers — system like suppressing results that are noise and instead saying "you need at least X people before you can make an informed decision". For Charlie's comments on all campaigns starting small, maybe means you would need to start campaigns slightly larger, as long as we can tell them reasonable sizes for how big they need to start for visibility

Erik: I think this is about our experiment platform with power tests, can you detect an effect. If there is more underlying variance, then relative noise doesn't change your decision much, but does make decision harder. Absent noise, you're making bad decisions all the time anyway! Noise doesn't make it worse. If you have good views of underlying distrib, then noise mixes it up. So the inference question is hugely complicated. But lots of users of this system don't use it to run test/control experiments. Without that, you can't really get these types of confidence intervals around conversion rates so that we could show the interval getting a little wider. I plan to add "this is the observed value, here is the confidence interval of what you would have seen without the noise", but *not* a confidence interval around true conversion rate.

Ben: There's always a cold start, "which creative is working best", might be good to test "You are running two creatives, how long do you need to run then before you can be confident which of them is better?" I've wanted this forever — people make decisions too soon based on not enough data all the time! Maybe simulate "I have two creatives with actual underlying conversion rate, I have some data on conversions, how long would you need to run to get certainty and how much longer does it get as a function of epsilon?"

Erik: There is a difference between when noise makes you make the platonic-ideal *wrong* decision, vs noise making you make a *random* decision. I don't think we should try to say "this is the right decision you *should* be making", people have lots of different motivations

Ben: Maybe we shouldn't chase the goal of letting people keep making irrational decisions? FB runs experiments, I'm sure many other people do too. Our stats team gets involved, says "this is how long you need to run to measure an effect of size X". Our experiment teams would be very curious to know how the noise affects that needed duration.

Erik: Gets back to the goal of what we're trying to measure: It's very easy for us with experimentation platforms to say "this is the way you should do causal inference", but I don't think most people make decisions based on A/B experiments, so we can't reasonably build a tool that pretends lots of people use that environment

Rachel: Is this predictive in nature? Is the data here based on historical data, real time adjustment?

Erik: Right now, it generates a conversion rate from a beta distribution by inducing variance, then simulates conversions from a binomial distribution from that conversion rate, then it repeats that to get results and maybe adds noise. No real prediction, just a simulation. We seed the beta distribution with a little bit of historical data from Criteo, but that was just to give us a starting point.

Charlie (Responding to Ben, ramp up budgets with broken config): Follow up on Ben and cold-start: If it's useful to test low-volume scenarios, there's the concern of an advertiser ramping up a campaign or new platform, and

they mis-configure their campaign or have some configuration failure. If the thing is broken and would always be producing zeros, we don't want to add noise and have them not realize their true data is all zeros! Want to distinguish zero from "you have some data but too small to be useful". Don't want to ramp up your budget based entirely on false positives on top of broken all-zeros measurement implementation.

Erik: Easy addition, we can just add a zero, interesting risk to think about. Maybe number of impressions for decision point is a valuable thing for the user to tell us. (still tricky question about conversion rate, but can hand-wave that for now)

Brian: We have some implicit assumptions about how controllable decisions really are. At the start of a campaign, you throw stuff against the wall and see what sticks, then keep making adjustments. Noise that we're introducing into the system amounts to giving people less clarity about where they are in their campaign as it's developing, which leads to bad decision-making. Creating much greater delay between time when something happens and time when you respond to it. Consider what Charlie suggested: As campaigns are developing, you get very general metrics about KPIs (right vs wrong direction), then get more detailed information once they have developed. That suggests machine mad to be making decisions and understand people's intent.

Erik: From API design point of view, makes sense that there are different questions you want to ask, different queries you want to run at different times, and maybe we can offer an efficient way to check that, good use of budget. Early, "is this making conversion?", then later other questions. I haven't heard any of that when it comes to API design! I'm hesitant to over-complicate this tool for some use cases and flows that hasn't even been suggested yet in these APIs.

Brian: It's a matter of the decision points when people interact. I'm suggesting that your tool should focus on the various decision points where people interact with the config of their campaign.

Erik: Charlie, correct me if I'm wrong, but I haven't seen proposals for these different points.

Charlie: A little more context here: We are initiating a study at Google, to study these other privacy mechanisms, where the point is exactly th study the low-epsilon high-privacy regime, where existing techniques will break down. Right now we let people choose their own epsilon, up to a fairly generous number! If that were restricted, these use cases would start breaking down, and that's what we want to prepare for. Additive noise may be fine if epsilon is really high, not if epsilon is tight. The reason we haven't put these forward yet is investment, but we are putting effort now into these new mechanisms, looking at the campaign lifecycle problem. We don't need your study to use these mechanisms yet, but good to understand effect of even simple mechanisms on decision points. There's lots of low-hanging fruit, but probably complicated, if we added another query to the table that did something more fancy. I feel like your survey could be really valuable in motivating that work more

Erik: We want to run this study. For that to say something meaningful, we should be tightly controlled. Simulation may be useful for people to get their hands on what this looks like. "Is this useful?" Maybe we should separate these two things. Maybe one is a study and the other is a playground.

Charlie: goes to what I said earlier: Very supportive of doing study as an MVP and then doing extensions later.

Graham Mudd: We shouldn't pre-suppose platonic idea for how to make decisions here. I worry that you are doing that, though, because you're asking for an expected conversion rate, then showing us measured rates compared to that, and of course lots of people will do the simple thing of being happy if it's over the predicted value. I'm worried that what we'll come away with is top-like conversion counts that you've suggested by definition, then you get do great, but I don't think that's real. Super-simple fix: When would people just keep things the same , don't spend more or less. Second thing: Ask for types of decisions people are going to actually make, so asking "how many cuts do you need on the data? how often will you make a decision?" Make

it closer to real world, then you will see things fall apart, budget stretched too thin, more realistic about how people actually make decisions.

Erik: That was the conclusion I tried to make on Tuesday — we could simulate the simple decision, we won't learn much; maybe the "no change" or "maintain" option might help us see something new and human, also maybe adding uncertainty bounds. Issue 28 on the repo is all about whether we should divide this by number of days, etc, please engage there, happy to think about variations.

Wendell: Repeat me and echo Graham: the question is what Ad Ops people will do in the face of uncertainty, it's a training tool, you should be able to say "x% of participants did Y, we expected them to do Z, please train you staff to do this not that". The cold-start problem is a great buzzword that you stumbled across, here's what we do: we go find the highest-comped person, take their campaign, add a little sizzle, re-up it. Your tool can help simulate that situation. This is an experimental what-if tool — anything you can solve with a math tool, please do it elsewhere. Anything about how my job is going to change, put it here.

Luke: Re being able to change questions mid-campaign: great to hear from people what they want to measure a different stages of campaigns. We want to be able to support those questions, with the minimal amount of noise that is necessary to protect privacy. We can change the questions being asked of the campaign to get more bins or fewer bins based on your volume — need to figure out how to do re-binning at various stages, to know what's important to being able to answer questions. Getting that feedback is hard.

Rachel: Q to Wendell: Are you talking about behavior modeling?

Wendell: yes. If a human is in the loop, then this is a human training activity, investigating decision-making with uncertainly. If you can do a math proof in an arXiv paper, do that and convince the math people. Uncertainly with machine in front of you is convincing for actual ads people

Erik: Assuming we get to MVP, we can certainly try to recruit folks who use our tools, but that's not the full picture. We want help recruiting, want people to volunteer to try it. Comment on the issue please. Want other folks to get testers in for the official, measured version of responses.

Brian: Would be helpful if we tried to categorize different types of ad techs — very large campaign different from local seller trying to generate business for their restaurant

Aram: Would be good to distinguish people there — getting buyer behavior into your data set is very useful no matter what. Useful to understand the scale of people decision-making, not for accuracy but for understanding the scale of inaccuracy. Would be good to give some notes of how much time this will take, how much math effort they need to put in, people may be reluctant to invest time or do things they think will be hard. Set expectations for participants to get them on board. I'm happy to promote, with that information.

Erik: maybe a small working group of us to decide that question, how much we want to ask etc — I don't want to make all those decisions myself. Anyone willing to spend an hour in the next week working on those sorts of parameters?

Aram: Let's have an ad-hoc. Brian, can you join something?

Brian: Ping me. It occur to me that we're implicitly asking the question "are the reports from these systems going to be used by people, or primarily by machines?" I can't help but feel like we're making things very complicated for human intuition, so people are making large general decisions but not finer-detailed ones like the scenarios being described here.

Erik: Machines have been involved in these systems for a long time, but people continue to be involved also. I don't think this is going to replace all the humans. Goal is to know when this falls apart for human decision-making.

Luke: Feels like currently decisions are made by both computers and humans. I'm trying to design a data stream consumable by the current ad industry, which means should be useful to both.

Aram: Ad hoc coming up!

Update on Steel man of proposals.

#152, Notes from TPAC

Aram: Doc came out of conversation at TPAC + breakout session that wasn't hybrid. Wanted to bring people's attention to it. There were also some back and forth conversations here.

Ben C: Can give a recap. At TPAC we had this comparison doc on the 3 main proposals and how they differ on different dimensions. On device / off device is the biggest difference. On device folks should see if they can converge on a design. Some discussion there. Luke hosted an Ad-hoc PAM call. Would love to get and update from Charlie and Luke there. Discussion points include measuring single events. Currently on pause. Other point is the importance of supporting a use case where there are a large number of Ad Ids and could support a sparse histogram. Opened an issue (link?) for how this would work.

Charlie: Thanks for the recap. Would add to that, in terms of the discussion on the PAM ad hoc call, spent time discussing API surface. Maybe not the most exciting thing, but gets to the sparse/dense histogram and sizes of them. In PAM, invoking the API involves bandwidth proportional to the size of our histogram size. Probably want to design something generic so it can work in as many ways as possible. A variant of ARA could also be workable with PAM. Task is to figure out if we can design a generic API surface that can support a wide range of types of histograms. Dense vs sparse is one dimension, size is also important (even if dense.) Different backends will be different. Also matters on the ε. Small ε can't support large histograms, moderate ones could. Hoping we could make progress to motivate these. Took an action item out of that for a straw man to design a PAM backend with the ARA front end. Maybe something I could spend some time on in the next couple months.

Ben: Happy to take this other thing to the issues. Want to understand the use case with large number of Ad Ids. Probably best to take that as a follow up item.

Brian: Encourage Charlie to write that doc, I'd like to see it.

Charlie: Sounds good. Ben, I can reply to the issue. Still not sure I've fully formulated if this is a big issue or not. For ARA we have a worse cartesian product issue. Right now, it's a research question tied to the issue on Tuesday about getting metrics.

Luke: In order to understand, it would be great to get an articulation of "this is the question I want to answer." Ben did a good job factoring the space, but would be good to see if there are any other factors. Solving for that or showing it's not possible would be great.

Ben: When you think of the breakdown key, it should be a cross product of source and trigger information.

Erik: Schedule ad-hoc call?

Luke: Big take away is to figure out where the on-device proposals are similar, and where they are irreconcilable. Large key space is one where we may not be reconcilable. Felt like progress in the last couple months is here are the shared capabilities of these two things. Charlie's effort to put a PRIO backend on ARA, then we could consolidate.

Charlie: Have made some progress. Opened some issues on PAM repo. One is on the notion of the filtering id, to make the spam filtering better. I think Luke said it was a good idea.

Luke: Yes, I don't like spam.

Ben: If we can go update it, we could make some things better.

Erik: Let's start async.

Ben: Action item: for low hanging fruit, we can async update the doc, tag each other. For larger issues with homework, we can wait and schedule some ad hoc meetings down the road.

Aram: If you get to the point where you have updated a bunch of these, feel free to put an adhoc meeting on the calendar.

Ben: Two docs: one that was presented, one that we're working on.

- IPA-PAM-ARA Comparison and Tradeoffs
- Steelman version IPA-PAM-ARA Comparison and Tradeoffs
- Also links if folks want to follow the discussion see
    - Issues on PAM
    - Notes from PAM ad hoc meeting

Processing pull requests

Re: https://github.com/patcg/docs-and-reports/pulls
[Aram: Seeing this as timely to address now and, seeing as the baseline review is a more open conversation, I'm pulling this up]

Nick: Question in Slack: how are we processing pull requests? Sometimes they linger. Unless we have a process, it won't happen, or we won't know what the status is. Prefer editor model, where editors can update and then we get consensus to publish. Or we can have group consensus on everything.

Aram: I was probably unclear on this. Everything in the docs repo, it us using the editors model. Everything should be marked as draft, editors can resolve PRs at their own pacing. My intent was not to get full consensus. Could open an agenda item if something is contentious and folks want feedback.

Sean: Do want to prevent an editor to just insta-merge things. Could tag it “editor ready” and then wait 5 days.

Charlie: Separate concern: at least one of the docs is the “doc of general agreement”. This is one where we do want consensus and not the editors model. (This is how the doc describes itself.) Would ask for a very clear model. Started as consensus in meeting, then did 2 week waiting period. In practice I wasn’t 100% sure if I was allowed to merge.

Aram: Two things: Not sure how we’ve been tagged in this in the past, I’m sure I’ve missed some. Unclear what is or isn’t ready except for “call for consensus” tag.

Charlie: Part of the reason is that we come to some consensus in the meeting, we write it down and tag other people who were discussing, no one else looks at it. We need some sort of rule: do we wait for the next meeting? Do we wait a certain amount of time? I don’t know what the process should be, but there should be some firm process.

Aram: Tried to avoid these being in a meeting.

Michael: Saying the same as Charlie. Pull 46 is an example. Had a discussion in the F2F in June. wrote up results, got LGTM from Charlie and Martin, made a few changes, then nothing happened at the next meeting. Then added call for consensus. No idea how this is supposed to move forward.

Aram: Just assumed that we should discuss call for consensus.

Michael: Last time we discussed this, we said we’d close all the issues with “call for consensus” tags in 2 weeks. (from TPAC minutes, 202309 Private Advertising Technology CG Minutes)

Aram: Ok, I’m misremembering. Here’s what I suggest: Any document that’s not “dimensions with general agreement”, let’s remove “Call for consensus” and allow editor model. The rest of these are ones we said we’d merge, so let’s do that. Moving forward, let’s add these, add the “Call for consensus” label, then in 2 weeks it can be merged if there isn’t any objection. Does that make sense?

Brian: May be over thinking things: maybe two labels? One for the two weeks, one for meetings.

Aram: Great idea. I will do that right now.

Sean: Could just use “agenda +” if we want to talk about it.

Brian: Worried people will be confused, will not recognize that there is automatic resolution after 2 weeks.

Erik: Can we write this down?

Aram: Yes, we'll pin the issues. One document that's not under normal editor process, the rest are.

Nick: Thanks for a process and for the chairs writing it down.

Michael: Who has the ability to add/remove tags and merge?

Aram: Anyone who's taken up an editor role.

Michael: I'm not sure everyone can actually do that.

Aram: There's a subset of folks who've actively added to this and I'll add those people.

Sean: No concerns. Folks can ask.

Charlie: Quick comment: People might be confused by labels etc. Let's add descriptions and link out to the doc.

Review Baseline Requirements for Private Measurement

Scribe: Paul
https://github.com/patcg/meetings/issues/141

Erik T: This is closely related to the Steel Man discussion (looking at the details of the actual issue)

Aram: We had three specific issues we wanted to discuss

Erik T: We were discussing different third parties. Most of these users will delegate to third-party (ad-tech or measurement company) – a company to manage these aspects. Do we think about worst-case vs/ avg case in these scenarios (are there restrictions specific to third-parties) and how do we enable third-parties to engage with these APIs. Design-wise, do want want "site owners" to enable access and control access to these third-parties

Charlie: We can answer some of these in the affirmative. As an example: "Can third-parties use these APIs" – I think we have consensus across all of the current proposals that they all enable this.
I think there are only some bullets we can make progress on – we need to talk about the third-bullet

Erik T: If I'm remembering correctly, it wasn't the simple question of enabling third-parties, the question was is there a specific way to invoke the API as a third-party that isn't just invoking JS on the page. The concern is that if there were multiple 3rd parties and there was a finite budget, one participant could consume the entire budget and starve out the others.

Charlie: I agree that's an important question to answer. I want to make sure we answer the simpler question - is it the intention of the API to allow for delegation? Should we draw stronger rules around who can invoke

portions of the API (eg: how PCM drew strong lines here) – so we should talk about “in what ways could calls/interaction be delegated?”

Luke: We talk a lot about delegation at Apple, all of our proposals will enable delegation.

Ben C: An additional dimension to consider is how difficult is the delegation action/mechanism? Consider PAM vs other approaches and if delegates need to be predetermined ahead of time (or if it happens ad-hoc) – “How should delegation happen?” is a question worth tackling

Brian: What is the interaction model for these APIs (including the parties and how they interact)? I assume we want model that the 1st party is control of all access and delegates including what they can/can’t do.

Aram: However the delegation is designed, allowing JS on the page (or using that as the main mechanism) is almost certainly not the path forward – this is already an issue today and the industry is rapidly moving away from it. It would be much more preferable to eliminate third-party JS on the page, and even better if we can avoid a network request all together.

Charlie (DOS / security perspective). On-device vs. off-device. Responding to Brian about the site author having control: Typically when we’re designing web platform APIs we design around the same-origin policy. When we talk about a scarce resource (eg a budget) and this is a “world writable storage/resource”, this introduces a violation of the same-origin policy – we need to look at the attacks that are possible (DoS over the resource). Any time there’s this challenge (supporting multiple 3rd parties), we’ll need to grapple with the possible security issues.
The secure delegation in IPA is one approach, but we could also take other directions (I know this group has looked at sub-budgets before). There are other options in the different proposals. We need to be mindful of the possibility of accidental adversaries, even given the best intentions around control.
We had lessons-learned in ARA related to fine-grained controls, but that level of control proved to be brittle and had unexpected results (including emergent attack vectors). In our current deployment some of these attack vectors are possible (we’ve controlled them with rate-limits), but it’s quite difficult to manage regardless.

Daniel: One aspect of the on-device that is a challenge comes from unauthenticated agents working on-device and taking fraud actions / spoofing.

Charlie: We’ve looked at that specifically in ARA. Only privacy budgets are the thing shared across origins in ARA to limit the scope of influence and attack vectors for bad actors

Brian: It would be helpful to understand the types of challenges you ran into ARA and the lessons you’ve learned (and how that shaped the design and evolution of ARA).

Charlie: We can do a bit of a post-mortem here and share more broadly. We can do a write up as well.

Brian: If you just wanted to point to GH issues, that would be fine.
https://github.com/WICG/attribution-reporting-api/issues/519
https://github.com/WICG/attribution-reporting-api/issues/558

QUEUE

1.
2.
3.
4.
5. **QUEUE CLOSED**

Participants Session 1

1. Alex Cone (Google Privacy Sandbox)
2. Alex Koshelev (Meta)
3. Alex Whitworth (Pinterest)
4. Andy Leiserson (Mozilla Corporation, a wholly owned subsidiary of Mozilla Foundation)
5. Aram Zucker-Scharff (The Washington Post)
6. Ben Savage (Meta)
7. Benjamin Case (Meta)
8. Brian May (dstillery)
9. Charlie Harrison (Google Chrome)
10. Chris Needham (BBC)
11. Daniel Masny (Meta)
12. David Dabbs (Epsilon)
13. Elias Selman (Criteo)
14. Erik Taubeneck (Meta)
15. Garrett Johnson (Boston University)
16. Graham Mudd (Anonym)
17. Hobert Bush (Mozilla Corporation, a wholly owned subsidiary of Mozilla Foundation)
18. Joey Knightbrook (Snap)
19. Kyle Hogan (MIT)
20. Lisa Markou (Ford)
21. Mariana Raykova (Google)
22. Martin Thomson (Mozilla Corporation, a wholly owned subsidiary of Mozilla Foundation)
23. Maxime Vono (Criteo)
24. Michael Kleber (Google Chrome)
25. Miguel Morales (IAB Tech Lab)
26. Nick Doty (CDT)
27. Paul deGrandis (Kevel)
28. Phillipp Schoppmann (Google)
29. Richa Jain (Meta)
30. Sam Weiler (W3C)
31. Simon Harris (DPG Media)
32. Taiwo Idowu (Google Chrome)
33. Tammy Greasby (Anonym)
34. Thomas Prieur (Criteo)
35. Wendell Baker (Yahoo)
36. Yuyan Lei (The Washington Post)
37. David Pham (The Washington Post)

Participants Session 2

1. Sean Turner (sn3rd)
2. Aram Zucker-Scharff (The Washington Post)
3. Brian May (dstillery)
4. Wendell Baker (Yahoo)
5. Charlie Harrison (Google Chrome)
6. Paul deGrandis (Kevel)
7. Daniel Masny (Meta)
8. Michael Kleber (Google Chrome)
9. Yuyan Lei (The Washington Post)
10. Kyle Hogan (MIT)
11. Ben Savage (Meta)
12. Martin Thomson (moz://a)
13. Taiwo Idowu (Google Chrome)
14. Sam Weiler (W3C)
15. Erik Taubeneck (Meta)
16. Lisa Markou (Ford)
17. Joey Knightbrook (Snap)
18. Tammy Greasby (Anonym)
19. Hobert Bush (Mozilla Corporation, a wholly owned subsidiary of Mozilla Foundation)
20. Benjamin Case (Meta)
21. Thomas Prieur (Criteo)
22. Alexandre Nderagakura (Not affiliated)
23. Luke Winstrom (Apple)
24. Graham Mudd (Anonym)
25. Andy Leiserson (Mozilla)
26. Mariana Raykova (Google)
27. Kazuhiro Hoya (Fuji Television)
28. Nick Doty (CDT)
29. David Pham (The Washington Post)
30. Rachel Yager (Evangelist and Chapter Head for W3C)
31. Richa Jain (Meta)
32. Phillipp Schoppmann (Google)
33. Miguel Morales (IAB Tech Lab)
34.
35.
36.
37.
38.
39.
40.
41.
42.

Cursor Nature Reserve:

ZOOM Link:

https://w3c.zoom.us/j/82659868398?pwd=R2wyMlVzVGcwcmZJb1BpZmdDc2crUT09