1

Good morning everyone. Thank you for coming. My name is Steve Weis and I am currently a technology policy fellow at the Aspen Institute. I previously worked for Facebook and was involved in the story I'll talk about today.

We often see privacy breaches or vulnerabilities from the outside. Rarely do we see what companies do in response. As a community, this hinders our ability to learn from our mistakes.

Today, I'm going to tell the story of one of those responses from the company perspective. While I am no longer an employee and can speak freely, I do have to generalize some specific details that are under NDA.

My hope is that as privacy advocates, you will better understand the competing interests at play and how they are balanced. I'll also make some specific calls to action that could help future companies in their own responses.

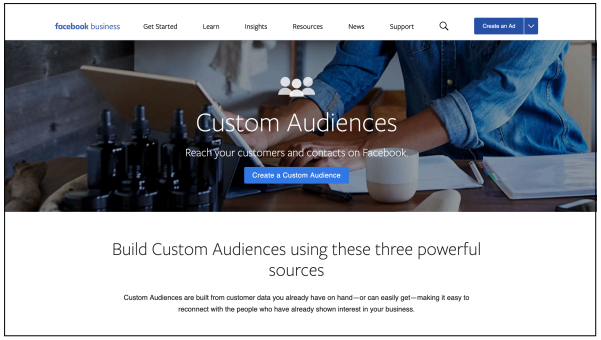
2

The story begins in December 2017 when a group of researchers from Northeastern University, France, and Germany privately disclosed a concerning privacy leak in Facebook's Custom Audience system.

Giridhari Venkatadri was the lead author of this work and has been one of the key people working in this field, along with his advisor Alan Mislove.

A month later, after giving Facebook time for an initial mitigation, the bug was disclosed publicly via an article in Wired.

I'd like to emphasize that the researchers in this case exemplified responsible and considerate disclosure, which kept the end user's interest in mind.

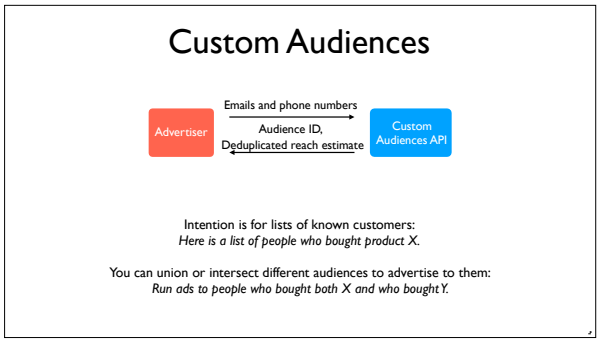


3

For background, Custom Audiences is a feature that allows advertisers to upload a list of names, email addresses, phone numbers, or other identifiers, and advertise to any that are Facebook users.

This is a very nice feature for small organizations. You can imagine a local shop with an email list that may want to run ads to past customers who are near their store. They can't afford broad campaigns or general advertising. But ads toward a small, specific group lets them reach a relevant audience for less money.

My premise is that we want to save this functionality. If you don't like advertising in general or especially targeted advertising, I'd ask that you suspend your disbelief for this talk.



4

How does custom audiences work? In this diagram, the advertiser sends a list of, say, phone numbers and gets back an audience identifier. They can later use that ID to target those users with ads.

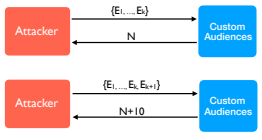
The API also returns a deduplicated reach estimate. If you upload a list of 10,000 email addresses and phone numbers, some of them might match the same user. Facebook will eliminate duplicates and return approximately the number who have accounts. Keep this in mind.

Critically, the advertiser doesn't know which users on Facebook match the identifiers they provided. They are just able to run ads to the audience in aggregate.

Deduplicated reach estimates are important because advertisers don't want to think that they are advertising to a much bigger audience than they actually are.

Facebook was burned in the past for overestimating video audience by accident, so is very sensitive on accuracy.

Finding a Threshold Audience



Threshold Audience A = $\{E_1, ..., E_k\}$

Adding one more real email or phone number will bump the estimate

5

Now let's look at the building blocks for the privacy leakage.

First is a notion of a Threshold audience. The Northeastern team noticed that the reach estimate would be rounded to the nearest 10, 100, or 1000. This was deterministic and reproducible.

What they would do is create an audience that was just on the threshold of tipping over that line. The addition of a single identifier would trip the estimate to the next quanta — for example, from 20 to 30.

Right off this gives you an oracle to tell if an identifier is associated with any Facebook account. That is already an information leak since a "Does this phone number have an account?" API doesn't exist elsewhere.

Exploiting Deduplication

Setup

1. Upload a list $A_{d,k}$ of every phone number with digit *d* in position *k*.
   Examples:
   $A_{5,1}$={*5000000, 5000001, … 5999999*}
   $A_{7,2}$={*0700000, 0700001, … 9799999*}

2. Pad each list $A_{d,k}$ with fake accounts to be a threshold audience.

Attack

1. Add a real target email to each $A_{d,k}$ and see if the threshold trips.

2. If it does not trip, that email's phone number has value d in position k.

6

For a local number, we'd create 70 audiences with 1M entries each. For each of these, they add on some fake accounts to pad them up to a threshold. The addition of a single real user who is not already in the list will bump the estimate.

Assume we check that an email address is a facebook user. We know how to do that with a threshold audience. For the attack, they will do just that: Add a target email to each threshold audience and see if the threshold trips. If it does, that

means it was not deduplicated.

If the threshold does not trip, we know it was deduplicated. That tells us the person must intersect with a phone number with d in the position k. So we learn a digit of their phone number.
By doing up to 70 of these checks, we'll leak every digit. Thus we learn what phone number an email address registered with.

What did they do in response?

7

This is where the public story mostly ends. Wired article runs, Facebook does something, and you never heard about it again.

What did Facebook actually do?

8

This is the Glomar Exlplorer, which was a ship owned by the CIA and the origin of the famous "can neither confirm nor deny" response.

I'm not an employee anymore and will neither confirm nor deny what the responses might be.

What I can do, is run through a lot of options that a company like Facebook might think about in response.

Non-Solutions

- "Just disable audience size estimates"

  Advertisers need to know how about how many people to expect to reach.

- "Just disable deduplication"

  Overstating audience size is worse for advertisers; accuracy is critical.

9

First, let's consider some non-solutions — assuming you want to save reach estimates.

Facebook is not going to straight up disable audience size estimates across the board. Advertisers need to get some sense of how many people are in an audience before they try running ads for it.

Second, they can't disable deduplication as was recommended by the authors. Again, this could blow accuracy since you will be double-counting users or worse. Facebook might get sued for something like that.

You could try to tweak the UI to say "Actual audiences might be 1/3 the size as represented" but then that is going to baffle users.

Disabling Multi-PII Reach Estimates

As noted in our advertising principles, we're constantly looking to improve the quality and security of our advertising platforms. After identifying a technical issue with reach estimation for Custom Audiences that could potentially allow misuse of the functionality, we're temporarily removing the ability to see audience sizes or potential reach estimates for newly created or edited Custom Audiences in Ads Manager and Power Editor, followed by the Ads API by December 22, 2017.

As of December 22, 2017, the endpoints below no longer support targeting specifications that contain the following custom audience objects (under either exclusions or inclusions), offline conversions API and Facebook Analytics:

- Custom Audiences from a customer file with **more than 1 key (multikey matching)**
- Data file Custom Audience (subtype = CUSTOM). The update does not apply to targeting specifications that combine a Data file Custom Audience with the same keys; for example, all custom audiences in the specs contain are based on emails.

10

The stopgap fix that Facebook publicly talked about was a combination of these two:
- Disabled reach estimates if multiple forms of PII are in an audience

The way a company would make that decision is to look at how many audiences fit that criteria and figure out how much ad revenue it would impact.

I will say that even this niche change results in hundreds of customers asking why their reach estimates no longer work.

This was intended as a short-term fix to buy time for a longer-term solution.

Traditional Mitigations

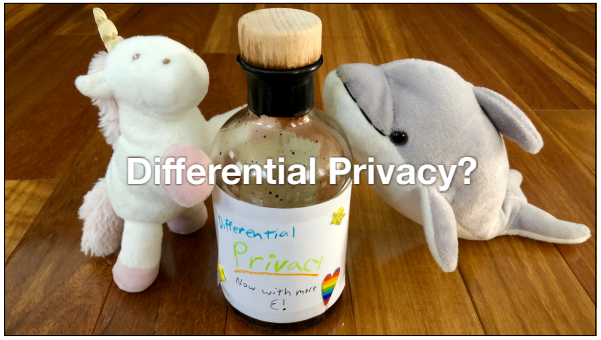
- Anomaly Detection:
  - *Legitimate people do lots of weird things.*
  - *This attack is easy to spread across many accounts.*
  - *You can create randomized lists of phone numbers to do it.*
- Rate Limiting:
  - *Legitimate people create lots of custom audiences.*
  - *This attack is easy to spread across many accounts.*
- Gating Features based on Ad Spend:
  - *Custom audiences is popular with small businesses with low ad spend.*

11

Now this particular attack involves creating a lot of odd audiences and making a lot of queries. Anomaly detection might help. In this case it's a challenge, because people use custom audience in weird and creative ways. It's also very easy to mask this attack by spreading it across many accounts or creating real looking audiences.

The attack also requires some number of queries to succeed. You could rate limit this to slow the attack down or cap the total number of queries. It turns out people create a lot of custom audiences. There isn't a reasonable rate limit that works across the board.

One idea might be to tune rate limits based on trust, where ad spend or longevity is a proxy. A challenge here is that this feature is popular with the small business, entry-level advertiser because as I mentioned, it is very useful to them.



12

What about this magic thing called differential privacy?

Differential privacy is a definition that parameterizes how much information is leaked with a given query. You can design different privacy mechanisms which can provide that parameter.

For example, adding noise to a result.

13

I can neither confirm nor deny that Facebook is using differential privacy for its reach estimates.

April 29, 2019

First Grants Announced for Independent Research on Social Media's Impact on Democracy Using Facebook Data

two-factor authentication and a VPN. In addition to building a custom infrastructure, we're also testing the application of differential privacy, which adds statistical noise to raw data sets to make sure an individual can't be re-identified without affecting the reliability of the results. It also limits the number of queries a researcher can run, which ensures the system cannot be repeatedly queried to circumvent privacy measures. We hope that this testing will lead to other benefits by letting us unlock more data sets to the research community safely and securely.

Source: https://newsroom.fb.com/news/2019/04/election-research-grants/

14

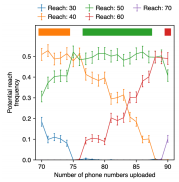
However, Facebook has announced that it is in-fact using differential privacy in other places.

Giridhari Venkatadri*, Elena Lucherini, Piotr Sapiezynski, and Alan Mislove

Investigating sources of PII used in Facebook's targeted advertising

3.2.2 Properties of noisy estimates

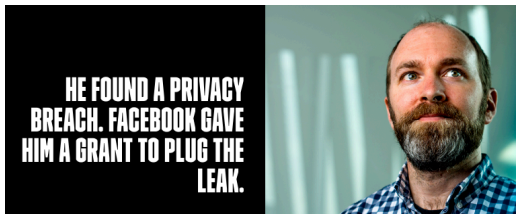
Since Facebook appears to be using noise to perturb the potential reach estimates, we move on to study how the noise is seeded, and to characterize the relationship of the noisy estimates corresponding to a given custom audience with the true value.



15

Also, Giri and Alan Mislove also just published a followup paper that observed perturbations of reach estimates.

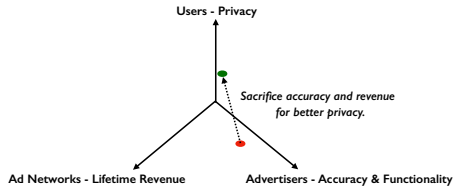
You can come to your own conclusions.



16

Another part of the response was the interaction with the researchers. In this case, Facebook paid out bug bounties and eventually gave a larger grant to the researchers for work on privacy-preserving aggregated statistics.

Balancing the Tradeoffs



17

Throughout this process, there was tension between the three main stakeholders: Users, advertisers, and the ad network like Facebook. They could have just axed the feature and written off the revenue and utility. That would be very hard to justify based on the risk. Instead, they had to tune how much to dial back revenue and accuracy in favor of privacy.

This is where easily quantifiable risk metrics are helpful : How expensive is the

attack? How long does it take? What types of data are leaked? What is the liability? One thing researchers can do is try to include their upfront and marginal costs in their publications. This can help convince decisions makers how easy or hard an attack is.

Epilogue:
Inferring Attributes


Exclusive: Facebook will no longer show audience reach estimates for Custom Audiences after vulnerability detected

Researchers were able to infer attributes of individuals using the tools.

Ginny Marvin on March 23, 2018 at 5:32 pm

Facebook Bug Bounty
March 23, 2018 ·

UPDATE JULY 2, 2019: Since suspending this feature last year, we've been working with researchers to improve the security of our custom audiences reach estimate feature. We focused on three key areas to curb potential misuse: privacy, protection and new usage restrictions. With these updates now in place, we are reinstating custom audience reach estimates. We continue to be grateful to the researchers who identified the bug and for working with us to fix it.

18

An epilogue is that the researchers found another leak of advertising attributes through the same mechanism. An Attribute in this case is basically a boolean about you: Interested in dogs, interested in cats, etc.

The challenge here is it's a single bit of information. Any throttling or probabilistic method is hard to use without degrading the feature. Ultimately, they had to decide which attributes were too sensitive to support.

Deciding what is a sensitive is full of nuance. Interested in Cancer Charities is not sensitive. Interested in Cancer Therapies is. It also varies greatly across cultures. Something innocuous in the US might get you killed elsewhere. Ultimately, Facebook disallowed reach estimates of audiences based on attributes. They just re-enabled this a month ago. I have no idea what the protections are.

Differential Privacy Calls to Action

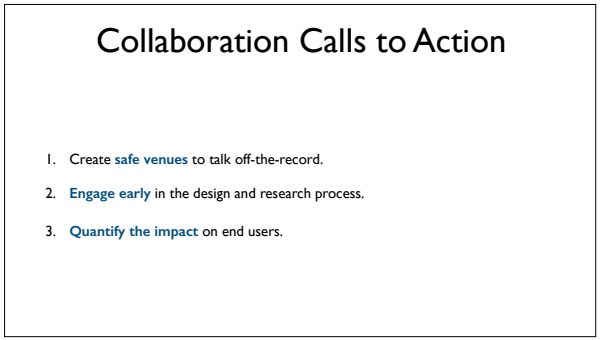
1. **More tools and libraries:**
   - Diffprivlib: https://github.com/IBM/differential-privacy-library
   - SQL query differential privacy: https://github.com/uber/sql-differential-privacy
2. **Case studies with specific privacy mechanisms.**
3. **Standardize parameters for practice.**
4. Think about **large-scale, dynamic, adversarial data.**

19

I'm going to close with a few calls to action. First, in the differential privacy field it's really not ready for prime time. We need more drop-in libraries and tools that non-experts can use. Having real world case studies like the US Census would help, so that people know which mechanisms are good for which cases. Same goes for practical recommendations of privacy parameters.

Finally, for researchers, I encourage you to think bigger, think dynamically, and

think adversarially. Much of the intro literature talks about static databases with a fixed set of queries. Something like custom audiences is a constantly changing data set, with data provided by adversaries themselves. Some of the privacy mechanisms in the literature that might help require maintaining a significant amount of state when you are talking that large of scale.



20

Finally, I'd like to put out some calls to action for collaboration. Companies are rightfully afraid to talk. They need to be precise in any public statement, otherwise can get sued and fined if they are found to be misrepresenting something. I'd love to see more off-the-record events where people can talk without attribution.

I'd also encourage both researchers and companies to engage early. Companies should pull in privacy advocates early to explain the issues, while researchers should give them a heads up on vulnerabilities early. I think this helps mutual understanding and speeds up responses. Both sides are taking a risk here.

Finally, I'd encourage researchers to try to quantify their impact on users. This is basically to give internal privacy advocates the material they will need to explain the risk. If you can quantify that something costs this much and takes this long, it is much easier to consider the tradeoffs.

21

That's I have and I'd like to thank you all for listening. I'm happy to open it up to questions now.