FTC PrivacyCon 2018
February 28, 2018
Segment 1
Transcript

KRISTEN ANDERSON: OK, everyone. We're going to get started, so please take your seats. Good morning, I'm Kristen Anderson. I'm an attorney in the Division of Privacy and Identity Protection, and I'm happy to welcome you all to PrivacyCon 2018. This is our annual conference, that highlights the latest in privacy and data security research.

Today, you'll hear from 20 researchers, presenting original work selected from more than 50 outstanding submissions. After each group of five presentations, we'll take a half hour to delve a little deeper, ask about the implications of their findings, and take some of your questions. And during the lunch break, in the conference rooms across the hall, we'll be hosting a student poster session, and government organizations that fund privacy and data security research.

Before we get started with our substantive program, I just need to address some administrative details. Please silence any mobile phones and other electronic devices. If you must use them during the conference, please be respectful of speakers and your fellow audience members.

Please be aware that if you leave the Constitution Building for any reason during the conference, you'll have to go back through security screening again. So please bear that in mind and plan ahead, especially if you're participating in the session.

Most of you received a lanyard with a plastic FTC event security badge. We reuse those for multiple events. So when you leave for the day, please return that badge to security on your way out. If an emergency occurs that requires you to leave the conference center but remain in the building, follow the instructions provided over the building PA system.

If an emergency occurs that requires the evacuation of the building, an alarm will sound. Everyone should leave the building in an orderly manner, through the main 7th Street exit. After leaving the building, you'll turn left and proceed down 7th Street across E Street to the FTC emergency assembly area. Remain in that area until instructed to return to the building.

If you notice any suspicious activity, please alert building security. Please be advised that this event may be photographed, webcast, or recorded. By participating in this event, you're agreeing that your image and anything you say or submit may be posted indefinitely at FTC.gov, or on one of the Commission's publicly available social media sites.

Please take your seats rather than standing. And please don't place belongings on the seat next to you. I know there's a bit of a line outside at security, so we do expect a full crowd today. Question cards are available in the hallway on the information table immediately outside of the conference room.

Christine, who I'm not sure if she's here or there-- right there-- will be available to collect your question cards. She also has question cards if you'd like one. Or, if you have a question that you'd like to submit, please raise your hand and she'll come get it.

For those of you participating by webcast, you can tweet your questions to @FTC using the hashtag, PrivacyCon18, or post it to the FTC's Facebook page in the Workshop Status thread. Please understand that we may not be able to get to all questions. The restrooms are located in the hallway just outside the auditorium.

And with all that out of the way, it's now my pleasure to introduce Acting Chairman Ohlhausen to provide some welcoming remarks.

[APPLAUSE]

MAUREEN OHLHAUSEN: Well, thank you. Good morning everyone, and welcome to PrivacyCon 2018. This week has been a whirlwind at the FTC. On Monday, we won a big case in the Ninth Circuit, which confirmed that our jurisdiction can reach the non-common-carrier actions of internet service providers.

Yesterday, we announced that PayPal settled FTC charges that Venmo misled consumers about the ability to withdraw funds and to manage the privacy settings. And today, we have not just PrivacyCon, but also a big relief that I'll talk about shortly.

And it's only Wednesday. Tomorrow, we'll have another big rollout, so keep your eyes peeled. Now, in such busy times, PrivacyCon couldn't happen without a talented team. And although there were too many people for me to thank them all individually, I would like to recognize Kristen Anderson and the whole Bureau of Consumer Protection team. Tim Daniels and the whole Bureau of Economics team-- our amazing Events Coordinator, Crystal Peters-- and the media team, and paralegals.

Now, this is the FTC's third-annual PrivacyCon. And privacy is a fast-moving field. And much has changed since the last PrivacyCon in January of 2017. Now I'm going to hit on three main points today. I'll summarize our work in privacy and data security in the past year, particularly our focus on the economics of privacy.

And building on that focus, I'll summarize my key takeaways from our Informational Injury workshop, which we had just last December. And finally, I'll talk about how PrivacyCon is a fitting capstone to my year as the Acting Chairman of the FTC.

So, first, the FTC has been extremely active in privacy and data security work over this past year. We've brought many cases-- and many important cases against companies like Uber, and Lenovo, and VTech. And just this week, as I mentioned, Venmo, among others.

We brought our first three actions to enforce the EU-US Privacy Shield Agreement. We are investigating the Equifax data breach.

And similarly, we've been active in our policy work. We held workshops and issued staff perspectives on the privacy and data security implications of connected cars, and of education technology. And of course there was our informational injury workshop, which I'll discuss in more detail shortly.

But first, the latest big news. Just this morning, we released a report on mobile security updates, that explores how smartphones and tablets receive patches for vulnerabilities discovered in their operating system software. We base this on information gathered using our 6b authority, which requires companies to furnish answers to specific questions.

Now, following a number of press reports about delays or lapses in mobile phone patching, we sought companies' information about how and when they deploy software updates to their devices. Now, this report is full of useful information about how security updates are deployed to mobile devices-- and the various roles that manufacturers, and carriers, and consumers play in a successful update.

We studied company responses to assess how fast they roll out updates, what factors affect the number and the speed of updates-- and how price, popularity, and age of devices affect manufacturers' decisions to update them.

And the report contains key lessons learned, and offers recommendations for government and industry. For example, it suggests steps that manufacturers should consider taking to deliver security updates faster to consumers. And it suggests ways that government, and industry, and advocacy groups can work together to help consumers understand the importance of security updates, and their role in the process.

Our Bureau of Economics also supplied a lengthy appendix with analyzes that underpin the findings of the report. And it's an in-depth, sophisticated report. And I applaud the dedicated staff.

And let me call out Lisa Jillson and Devesh Raval, and Nathan [INAUDIBLE], who worked so diligently to produce it. And I hope you will all set aside some time to read it. Now, as many of you know, over the past year I've sought to explore the benefits of applying an economic approach to privacy and data security issues. And today's lineup reflects that emphasis. And I look forward to hearing from our panelists who have pursued this type of analysis.

But what does an economic approach look like? It is decidedly not just about numbers, and measurements, and formula. That's mathematics. Now, many economists use math, but economics isn't simply about math. Economics is about real people making choices about how to use limited resources to get what they need and want through exchanges in the marketplace. Economists seek to discover general principles about those individual exchanges, and how in aggregate those exchanges affect society.

Thus, an economic approach to privacy means applying the tools of economic analysis to help understand how and why companies collect information-- what exchanges are taking place, and

the likely consequences of certain arrangements regarding private information. What makes information valuable? How valuable is it? Does the value depend on who holds it, and why?

Why, and how do people exchange information? How can information be put to its highest-value use. Does it matter who gets to decide the highest value use? And, how do markets and information develop? And, what are their flaws?

These are important economic questions. And they are key privacy questions, too. And I'm glad to say that, at the FTC, we've already been looking at privacy and data security using economic thinking.

Now, the most visible result of this work was our December Informational Injury workshop. And this full-day workshop sought to better identify the qualitatively different types of injury to consumers from privacy and data security incidents. It examined different definitions of informational injury, and when government intervention is warranted.

It explored frameworks for how we might approach quantitatively measuring such injuries, and estimate the risk of their occurrence. And it sought to understand better how consumers and businesses weigh these injuries and risks when evaluating the trade offs to sharing, collecting, storing, and using information.

Ultimately, the goal is to inform our case selection and enforcement choices going forward. And let me share a few of my key takeaways from the event. First, it is clear that privacy and data security incidents can and have caused injuries that do not involve solely financial loss.

The first panel discussed real stories about such injuries. These included medical identity theft affecting medical treatment, and doxing-- which is the public dissemination of private facts leading to extortion or stalking.

Now, this reminds me of the case we brought earlier this year against MyEx.com, a revenge porn website. The site urged users to get revenge by posting intimate pictures of others. And then, it would charge the pictured individuals for taking down the photos. And people who were featured on this site suffered real harm in addition to the money they paid to remove intimate images and personal information. Many lost jobs, or job opportunities-- or were threatened, stalked, and harassed.

The panel also talked about increased risk to health and safety that can arise from the revelation of people's real-time location information. Now, the takeaway is clear. Consumers can suffer injury from privacy and data security incidents, and that injury isn't limited to loss of money. And this is also consistent with the FTC's long-standing case law, and the Commission's deception and unfairness statements.

Second, although it was clear that injury is more than just financial loss, there was wide disagreement on the second panel about what else compromises privacy or data security injury. Now, all agreed that certain uses of sensitive data can cause injury, even if the harm was not solely financial.

However, some had more expansive views of injury-- including concepts such as breaching the boundary between one individual and another, revelation to others of something private, increased risk or likelihood of a future cost, or contravening a person's expectations without any benefit to them.

Now, some of these definitions of injury were consistent with previous FTC actions. But some would sweep in nearly all information collection. The extremely wide range of conflicting conceptions of injury demonstrates that defining injury is a key issue in the privacy debate.

Now, despite the experts' wide-ranging and conflicting definitions of injury, there was general agreement that government intervention ought to be tied to injury, whatever the definition. All panelists agreed that some injuries do not warrant government intervention, although they differed significantly on when intervention might be warranted-- again, largely based on how they defined injury. All panelists agreed that countervailing benefits have to be evaluated as well. And panelists also agreed that there were trade-offs to both ex-post and ex-ante interventions.

Now, the fourth topic of the day was measuring injury. And to paraphrase one of the panelists there were easy cases and there were hard cases. Unfortunately, most of them are hard cases. And research in this area is challenging. And while people say they care about privacy in the abstract, what they do when faced with actual choices is often very different.

This isn't necessarily because consumers don't know what they think. Instead, consumers balance a huge range of product and service dimensions when making privacy choices. And privacy is but one important dimension of a variety of products. Others include price, convenience, and quality, and other factors. And different individuals will make different trade-offs when faced with these many dimensions.

So in trying to measure injury there are many things one might consider-- the type of injury, the sensitivity of the data. The magnitude, the frequency, and the causal link to a particular firm or practice. And each of these raises significant challenges. And there is interesting work on each of these dimensions, but we certainly need more research.

And I also note that different types of measurement might be appropriate for different purposes. In the law enforcement space, for example, tools might differ depending on whether the goal is selecting cases, demonstrating liability, or calculating damages.

Overall, my key takeaway from the Measurement panel was that not everything that can be measured matters. And not everything that matters can be measured. But we ought to measure the things that we can, and think hard about how to objectively and consistently evaluate the things we cannot. After all, if we cannot measure or even estimate the injury we're trying to address, how can we tell if we are directing government action effectively?

Now, the Informational Injury workshop was another chapter in a conversation in research agenda that it's only getting started. Today's PrivacyCon continues that exciting conversation. And PrivacyCon has a couple of purposes, as I see it.

First, it helps the FTC stay up to date with novel, interesting research that can inform our privacy and data security missions. Second, PrivacyCon provides a forum for highlighting emerging issues and bringing them into the conversation. And today's program will fulfill both purposes.

The emphasis of previous PrivacyCons has been technological developments. And today's program continues that developing tradition. Panels one and four contain presentations describing how technology can exacerbate or alleviate privacy and data security risk.

As I've already mentioned, this year we also emphasize economic questions-- with 10 papers in panels two and three, that explore consumer perception and behavior, firm incentives, and market characteristics. And I think this cross-disciplinary approach is essential. Lawyers, technologists, and economists can learn a lot from each other.

And given the challenges consumers face on the privacy data security front, we need to take advantage of crosscutting research. Doing so will help consumers, industry, and the FTC better understand how to protect privacy and data security.

I'm also gratified that during the lunch period we will be able to showcase in our poster session the work of early career academics and students. And I'm pleased that attendees will have the chance to interact with representatives from government agencies that fund privacy and data security research, as well. Both the poster session and the funders will be in the conference rooms just across the hall-- that way-- during the lunch hour.

To summarize, it's been a big year in privacy and data security at the FTC. We've accomplished a lot for consumers, and we have a lot of work to do going forward. And PrivacyCon is an important part of that work. So I thank you all for being here today. And best of luck to the presenters. And I look forward to a productive day. Thank you.

[APPLAUSE]

DAN SALSBURG: Panelists on the first panel, please come up. Thank you, Acting Chairman Ohlhausen. Welcome to the first panel of PrivacyCon 2018. I'm Dan Salsburg. I'm the Acting Chief of OTech, the FTC's Office of Technology Research and Investigation.

Brace yourselves. For the next seven hours, you're going to hear 20 research presentations. To make this digestible, we've grouped the research into four sessions of five presentations each. As Kristen mentioned at the beginning, there will be a brief discussion period after the five presentations.

This first panel focuses on the collection, leakage, and exfiltration of private information. What are the privacy implications of opening an email, or clicking on a link in an email? How do browser extensions track users? What are the privacy risks of end-user programming-- so, if this, then that applets. How can session replay scripts, that follow a user's movements while visiting a website, impact that user's privacy? And, how could Facebook's advertising platform be used to obtain a user's PII? Prepare to find out.

I'm joined today by five researchers who have studied each of these issues. In the interest of time, I will not read their bios. They're available in the handouts today, and also on our website-- PrivacyCon website.

But to briefly introduce them-- on my left is Steve Englehardt. Steve is a PhD candidate in the Computer Science Department at Princeton. And Steve also interned with our office, OTech, last summer. To his left is Michael Weissbacher, who is a doctoral student in the College of Computer and Information Sciences at Northeastern.

Next to Michael is Milijana Surbatovich, who's a PhD student in the Electrical and Computer Engineering Department at Carnegie Mellon. Next to her is Gunes Acar, who is a post-doctoral research associate in Princeton's Center for Information Technology Policy. And finally, at the very end, is, Alan Mislove, who's an Associate Professor, Associate Dean, and Director of Undergraduate Programs at the College of Computer and Information Science at Northeastern.

To get things kicked off, we'll have Steve present his research on email tracking.

STEVE ENGLEHARDT: Thank you, Dan. So, today, I'm going to talk about email tracking. And when I say email tracking, it's not the kind of send tracking that you might think of when you hear that term-- like, being able to tell that someone opened an email, how many times they opened it, something like that.

But it's much different. It's looking at a marrying of web tracking with email tracking. And I'll show you the implications of that, and how that has far reaching effects in terms of user privacy online.

And so, I'm sure many of you have seen something like this, right? You open up an email, and you see a message at the top that says, remote content has been blocked for your privacy. And then, you see a bunch of images which weren't loaded. And instead, you see maybe the alt-text. Maybe you just see the outlines of those images.

And this really makes emails completely unreadable, right? You don't have no idea what this sender was trying to say. You don't know, say, what deals they were trying to send you, or so on.

In a more serious case, it might be a bank sending you a notice. And you might not know what the notice says, because you can't read it. So if you're like me before I actually went and did this research, you'll say, OK, I trust this sender. I know Bed Bath and Beyond. I'll open up the email. I'll allow the remote content to load. And so the question is, when that happens, what are the privacy implications?

And as I said, it's not just this send tracking that you see here. It's not the notion of, OK, well the company now knows I've opened this email a couple of times when I opened it. Instead, it's something much different. Emails are being tracked far beyond just the send tracking.

And I'm going to show you a specific example here with LivingSocial. So let's say we open up LivingSocial. We allow images to load. We allow remote content to load. The device you're

using ends up making requests, or contacting 24 different companies-- 20 of which can track you, if your email client or your web browser supports that. And 10 of which will receive an MD5 hash of your email address.

So if you're not sure what hashes of email addresses are-- I'll go into a little bit more later. You can think of it as an encoding of the email address with some with certain properties. I'll discuss that more later. But what you see here is a lot of familiar companies involved involved in web tracking. You make requests to domains from Google, and Oracle, and OpenX, and so on. You have data brokers like Acxiom involved.

And so the question is, why would they want to receive a hash of your email address? And, what could they do with the tracking information they receive? That's what I'm going explore in this talk. And that's what we explore more thoroughly in the paper.

And I want you to think of email tracking as basically web tracking without JavaScript. So just as you can use web tags to kind of follow a user around the web-- a small image tag that's invisible, you put it on every website-- you can do the same thing in emails, as long as remote content loads. And so I'm going to talk about what kind of tracking you can do when basically you just have those web tags available to you.

And so to measure this, we went and crawled a bunch of sites. We automatically signed up for mailing lists on those sites. We ended up receiving about 13,000 emails from 900 of the sites. And we measured tracking with our measurement tool, called Open WPM So with each email, we went and opened it up in this kind of simulated webmail client. And we saw which resources were loaded, and what kind of tracking was present.

And so I'm going to just give a few highlights from the paper. I'm not going to go into detail on all of our findings, but I encourage you to check it out. We will have a link at the end of the talk.

So one of the things we were kind of surprised by is, that many of the top web trackers are in kind of a large portion of emails. So DoubleClick, for example, was in almost a quarter of the emails that we received. And actually 85% of the emails embed some type of third party, with an average of five third parties per email.

So this was by no means an equal distribution. There were some emails, like I showed you with LivingSocial, that had 24 different companies. And there were some that had none. But more troubling is, we found a bunch of emails-- and from a bunch of senders-- that were leaking email address to third parties-- often intentionally.

And so you'll see some kind of plaintext leaks. Basically you can think of it as an image tag that has an email encoded in it. And you also see some of these hash leaks that I talked about before. That's things like MD5 SHA-1, SHA-256. It's different hashing algorithms, or different kind of encodings of the email.

And basically what a hash is, it's something that it's easy to compute. So it's easy to take an email address and compute the encoding. But it's very, very difficult to go in the reverse direction-- to

take only the encoding, and with no other information try to figure out what email it corresponds to.

And some of the things you can do with this is, you can actually correlate web tracking and email tracking. And I kind of show a diagram of it here. I'm not going to step through it. But we do step through it in the paper.

And the notion is, just like you can use web tags to kind of follow a user around the web, you could also use the same web tags in combination with email addresses-- leaked email addresses-- to connect web browsing and this kind of email-based tracking.

And so instead, I'm going to focus on this notion of people-based marketing-- essentially, marketing to a real person rather than marketing to a cookie. And I'm going to focus specifically on Live Intent, which was the largest receiver of email addresses in our study. So they received the most email addresses from the most number of emails.

And so if we dig into some of their marketing material, we see certain things. They actually have a blog post that says, email addresses are so much better for tracking users than cookies. They are deterministic. They're persistent.

So basically, if you receive an email address you know that, OK, this corresponds to one user. You know that that user is probably not going to change their address at any point. It's also cross device, right? Users will log into the same email on all their different devices-- their mobile phone, their desktop, their work computer.

And if you are receiving these tracking signals from each device, you can connect all the tracking for all of them together. And actually, when you dig into their privacy policy, they say that not only will they collect this online information, but they can use it to combine with online information.

And so you can imagine this as, you go to the store. As you're checking out, you give your email address to the clerk at the store. And maybe that ends up in a database somewhere. That could be kind of onboarding-- data

And so, what are the privacy implications? When you look into the statements that the companies make-- they say, well, we hash it. That deidentifies the information. And it does not permit your reidentification. And so let's look a little deeper into that claim.

So let's say that we have a tracking database that looks like this. We have an email hash, which is the identifier. So that's what a hash might look like-- kind of a random string. And then, we have a bunch of tracking data associated with it-- might be web visits, if it was connected to the user's web tracking. It might be the emails that they've opened. Or, it might be their purchases, like I said, at these point-of-sale data collection.

So if you want to run a deanonymization attack on this, you can just open up a terminal in your browser and enter that line. And the point I'm trying to make here is, that because we know the

email address, it's very easy to determine what the hash is-- because all you have to do is hash a known address.

And let's say you have a tracking database full of every American consumer. You can be pretty sure that at least one record will match whatever address you end up hashing. And that's because, like I said, going from an email address to a hash is easy. But going from a hash string to an email address is difficult if all you have is that hash string. But it's easy if you can guess the inputs to the hash function-- if you can guess possible email addresses.

So for example, take all the email addresses from the authors of this paper-- and this slide got a little messed up. And one of them ends up matching, say, this database record. And so you can actually do this at scale. Email addresses aren't secret. In format, they're pretty predictable. But even if you don't want to try to predict the format of them, there are tons of database leaks that someone could just go out and download a billion email addresses.

And if you still have some problems-- like, some of the addresses still aren't in those leaks-- you can just guess them. So we went and did kind of a back-of-the-envelope calculation here. And there are some papers that said there's about 4.5 billion email addresses in use.

And if we were able to guess one of those addresses every one in a million guesses that we made, we could generate that whole space for $75 very quickly.

So the point is, you shouldn't think of email addresses as secret. And thus, hashes are likely to be pretty easily reversible. And actually, past research that's done this on real databases has had a pretty high success rate-- 45% to 70% of emails.

And if you don't want to guess, you can just pay someone else to do it. So there are a bunch of services that offer email hash reversal as a service, for as cheap as $0.04 per email. And beyond that-- the one on the bottom of the slide there actually says they'll connect it with consumer data for just double the price. So if you want to get a bunch of data associated with that hash, you can do that as well.

And so I think the takeaway here is that the claim of email addresses being identified is, I think, really a weak claim, and something that needs to be looked further at. And that there is this industry being built around email address hash-based tracking. The claim is, these are the identified identifiers. You can get them from every one of the user's devices.

And it's so much better than your traditional tracking things, like cookies. And, that the line between email tracking and web tracking is pretty blurry. So you can check out the full paper at the link at the bottom of the slide. And you can check out some more of our work on Thinker. Thank you.

[APPLAUSE]

DAN SALSBURG: Thank you, Steve. Now, Mike will speak.

MICHAEL WEISSBACHER: Thanks. So I will talk about the X-ray detection of history leaking browser extensions. This was a joint work between Northeastern University and University College, London.

So before we talk about the privacy leaking and browser extensions-- [INAUDIBLE] browser extensions. So this is third-party code, which is added on top of a browser like Chrome or Firefox. So these are developed by different people.

And so what these extensions offer is, powerful APIs, who can access a lot of things-- which can be restricted for permissions. So they can modify the page as being currently visited. It can change requests and responses made by the browser. They can often access any of the pages which are being visited. And they have access to cookies. They can access history which has been browsed before installing the extension.

So if you don't know whether a browser extension is, usually if you look into the browser and you have icons at the top right, these are belonging to extensions. So they're ubiquitous, so the popular extensions have millions of users. There's thousands of them. And they're good.

So they enable us to do ad blocking, or sharing notes on websites. So this is good. But on the downside, we have these privacy implications coming from these [INAUDIBLE] extensions.

So this permission system, which I mentioned, it's very good to restrict some things. But to contain these privacy leaks, it is completely inadequate. So to enable an extension to leak old browsing history, it's enough to fall into the category "low alert" in the Chrome Extension Store. So this collection is sometimes mentioned in terms of service. So when you scroll to the bottom of the terms, sometimes they will say, we are collecting all your data when you browse. But sometimes, it is not.

So the question is really, can a user be expected to read the whole terms of service? And do they have to expect that they're being fully tracked when they install an application, which seems to be doing something unrelated?

And another thing is automated updates, which are good in general. It's good to keep the extensions automatically up to date. But the downside is, if it's being updated, and code is added to decode browsing history, you might be not aware of this.

And there is no tool that will tell you this extension is leaking. Or, there is no indicator in any of these extension stores that will tell you, this might be an issue for your privacy.

So to contrast this web tracking, which is well studied-- or more studied than the extension tracking-- this is quite different. So to be tracked on the website, the first step is that the website owner has to opt in, essentially, and install this tracking code.

So, often this happens. Also on the user side, there is an option to use an ad blocker, or a tracker blocker, which will somewhat reduce the impact on the user. So only both of these are true. So if there's an opt in and no opt out, the user can be tracked.

An extension is different. So, usually, installing an extension gives access to all websites-- so anything the user is visiting. And the opt in is implicit. So by installing this, tracking is active. And there's no opt-out option, other than installing these extensions. So there is no button one could click.

And so, the motivation for this project was, we found one library that was used in 42 extensions. And all extensions that were using this library, they were leaking all browsing history. And this was quite popular.

So 8 million users had this library through these 42 extensions. And after finding this, we documented this in a blog post. And we just put this online. And Google deleted all of these extensions in 24 hours. So they were gone from the store, which was encouraging, but there was no change in policy.

So it's not prohibited to upload extensions which are leaking all browsing history. So we did the next step. We did the honeypot probe.

So we took all these extensions that we could find, and ran them in isolation in a container. We provided unique URLs to each of these extensions, so we could identify them. And we are browsing our website in this container. But we also let it connect to the internet.

And when we saw third-parties connect to our domain on the public internet, then we knew that there has to be a leak. So what did this look like? On this plot, we have on the x-axis, time. The y-axis are unique extensions that were leaking.

And the blue dots are when we were executing the extension. And the Xs are when we received incoming connections. So if you want to take a look in more detail, it's in the paper.

But the results of this experiment was that, as incoming connections proved, it is not only being leaked, but also being used somewhat. So they were collecting information on the pages which are being visited. These extensions in the last slide were used by 3 million people.

And so, this excludes the extensions that we mentioned earlier. These connections often happen immediately after being executed. So when running these extensions and visiting pages, there is often going to be a second connection coming from somewhere else. But this was obvious, this is just a lower bound of leaks. So these extensions which we saw like this are only the ones that would immediately connect to our server.

So more may be leaking, and simply keep the data and process it later. And we also saw indicators of collaboration. So some of these domains were contacted by multiple of these trackers.

And there were also multiple servers connecting to a single extension domains. And based on this, we did a [INAUDIBLE] system. So this is X-ray. We wanted to detect these type of leaks automatically.

And we had two goals, really. We wanted to be robust about the detection. But both of how the data is being collected. So there's multiple ways to do this in the browser. And also, how the data is being exfiltrated. So it can be either obfuscated or encrypted using different protocols. We wanted to be competitive-- all of them. So we built two complimentary automated detection systems. But also, another system that will help an analyst do this manual analysis faster.

And the data which we used was network traffic, but also based on an instrumental browser, which we used. So we modified Chrome by rewriting C++ code. And we analyzed all extensions with more than 1,000 users. And these were over 10,000.

And so, how does network side looks like? So we used counter-factual analysis. And we based this on the properties of how these trackers work. So the modification to the size of history-- so we used this as a variable. And then, increased the amount of history which we expose to the extension. And the network has to change behavior if the extension is leaking effectively.

And so what we see in the plot are, four runs of a couple of extensions. So this is normalized to the first run, which is no bar here. And the right side is, benign extensions. And these are essentially behaving the same, regardless of the size of history.

On the left, we see extensions which are leaking. And they increase the amount of send data to these tracking domains, based on how much history we expose to them.

So our findings-- so we found over 10 million active users for these extensions which are leaking history. And we analyzed 10,691 extensions in total. Of those, we flagged 212. But we also had false detections. So we flagged 28 extensions which were not really leaking. So we had a false detection rate of 0.27%.

And we also found two novel ways of leaking-- ways of sharing this data which was not known before.

So to wrap this up-- so history leaks are a big issue in browser extensions. And these extension stores don't seem to be checking for this behavior, as far as we can tell.

However, it is possible to have robust detection techniques for this. And two possible remediations-- so, one, it would be good if these extension stores would check for this type of behavior, and analyze their extensions. But the only tangible thing we can really suggest is, that users delete extensions that they don't use, or seem suspicious. And there's a link to the full paper in the slides. Thank you.

[APPLAUSE]

DAN SALSBURG: Thanks, Michael. Next, Milijana will present her research on the security and privacy risks of if this, then that recipes.

MILIJANA SURBATOVICH: Thanks for the introduction, Dan. I'm going be talking about security and privacy flaws, and end-user internet of things programming.

So the internet of things is made up of smart devices, which are the wireless-enabled home appliances and the online services connected with them. And these smart devices introduce new security concerns.

Sometimes these happen through attackers, such as an incident called the Mirai Attack in late 2016, when attackers injected malware into thousands of these smart devices-- and used it to take down access to large portions of the internet.

Sometimes these security concerns happen accidentally, such as an incident a few weeks ago, when Strava released the public heat map of its data, and it revealed the locations and outlines of US military bases around the world.

But we're specifically looking at this service called If This, Then That-- or, IFTTT, which allows users to connect the behavior of these already potentially problematic IoT devices and online services together. And they do this through allowing users to create if, then rules such as, if my Fitbit logs 10,000 steps, then it will automatically post a tweet about that. Or, if I'm close to home, then it will automatically turn on my Nest thermostat.

And these rules are obviously very useful and convenient. But the question we're interested in asking is, whether these IFTTT applets can have harmful, or just unexpected side effects. So to go into the terminology a little bit before I continue, this if, then rule is now called an applet. Previously, it was called a recipe. If portion is called a trigger, then that portion is called an action.

So some examples of trigger events would be-- receiving a new email, reaching your step goal, being tagged in a photo. And some examples of action events would be-- creating a new status, setting the temperature, or uploading a public photo to some service.

Now, things can go wrong with these applets. Consider this one. If I take a photo with my iPhone, then automatically upload that as a public photo to Flickr. So you might want this rule if you're going on some touristy trip. You're going to be taking lots of nice photos, and you want them backed up automatically, and available for people to admire.

But it can have an unintended consequence. If on your trip you run into some visa issues, and you have to take a photo of your passport to send to some official. So then, now that's available on the web.

So we call this type of problem a secrecy problem. Applets can leak private information. Through this applet, the information-- in this case, your passport photo-- is going from being available just on your phone, available to you, to being available on the web-- visible by anyone.

That's not the only way things can go wrong though. Consider this applet. If I'm tagged in a photo, then automatically make a new Facebook status with that photo. So you might want this applet if you're going to some event, but you're not the one taking the photos. And you still want them available on your Facebook page, so people can see what you're up to.

But it can have an unintended side effect, if you call in sick to work to go to some party. And you're not taking photos, but your friend does and tags you in it. Your boss sees, and then gets mad at you for slacking off. So we call this type of problem an integrity problem. Applets can be triggered by untrusted or just unintended sources.

So in this case, you gave up direct control over what's appearing on your Facebook profile to that friend who tagged you in the photo, when perhaps you didn't want that. So what we want to do is, systematically analyze applets for potentially harmful or unexpected side effects. We did this through categorizing the secrecy and integrity levels of each triggering action.

And you can think of levels, in this case, as being similar to the top-secret, confidential classification labels that are used in the government. So once we came up with the labels, we want to define what makes a combination through the triggered action of a recipe unsafe. And then, we use this to analyze applets at scale.

So to define secrecy levels, where secrecy is referring to who can see or know that an event is taking place. We have private events, that only you would know about by default-- such as logging a new weight on your Fitbit.

We have restricted physical events, which refer to those that those nearby could see-- such as the temperature on a thermostat. If it's in your home, your flatmates could see it. Or if it's at work, your coworkers could see it.

We have events that take place close to you online, as it were. So events that your Facebook friends or Twitter followers could see-- such as you updating your status. And finally, we have events that are available to anyone-- visible to any one-- such as a new article on The New York Times. So once we came up with these levels, we connect them with arrows going upwards. And that makes our secrecy lattice.

If the flow from the trigger to action of an applet flows upwards through the lattice, then it's safe. It's not giving up any information, because it's going from a broader audience to a smaller one. On the other hand, if the flows from the trigger action goes down through the lattice-- if information is going from more restricted to less restricted, then it's potentially leaking sensitive information.

So I'll go through some examples-- this time, considering the labels. This recipe that we saw earlier-- if I take a photo with my iPhone that's a private event, because only you are seeing that photo, whereas uploading as a public photo to Flickr is obviously public.

Or for this recipe, if I have a new Strava activity-- so that's a jog or a run-- then it posts a tweet with the image of that path. The new Strava activity is private, depending on your Strava app settings. But posting a tweet with an image is restricted online, because your Twitter followers can see it. And this recipe is leaking your behavioral or daily routine data to your Twitter followers, who may misuse it.

The integrity levels are analogous to the secrecy. But integrity is referring to who can cause an event. So we have untrusted events that anyone could cause, such as a new top post on Reddit. We have restricted physical and restricted online events, which are those that nearby can cause-- such as asking Alexa in the house, or tagging you in a Facebook photo. And then we have trusted events, that only you should be able to cause, such as sending an email from your email account.

So again, we connect those with the arrows going upwards to make our lattice. And the trigger to action flowing up the lattice is safe, because it's allowing a trusted event to control less-trusted action. On the other hand, if the triggered action flows down through the lattice, then you're compromising the integrity of that trusted action by allowing it to be controlled by that untrusted source. So some examples of integrity violations would be this applet that we've already seen.

If I'm tagged in a photo that's restricted online, integrity for the trigger, because my Facebook friend is tagging me. A new Facebook status is a trusted event, such as is happening from my accounts. In this recipe, if I get a new attachment in my inbox, upload that file automatically to my OneDrive. A new attachment in my inbox is untrusted, because anyone can send me an email with my attachment. Because as we saw, emails are not secrets.

And then, uploading a file to OneDrive-- depending on who I've shared my OneDrive folder with-- if I haven't shared it with anyone that's a trusted event. And this recipe is making it easier to sync malware to anything that's connected to my OneDrive, because that attachment could have been malicious.

So once we came up with these labels and lattices, we used them to analyze the IFTTT data set of all the applets that existed in early 2016. So it had 200,000 total applets, and 20,000 unique, because people made duplicates. So we labeled the triggers and actions with the secrecy and integrity levels, and then ran queries to select what violated-- so, flowed down through the lattice.

What we found from these 20,000 applets are, that half of them did violate these rules. Around 2/3 of the violating rules had integrity problems, and half had secrecy. More specifically, we found that the most common type of secrecy violation went from private, to restricted online, to restricted physical.

So users are giving up information about themselves to a still somewhat restricted audience. An example of that would be, if I reach my daily step goal and upload that information to Facebook.

For integrity violations, the most common flowed down through the lattice-- end up at that trusted level-- with the most common being untrusted, all the way to trusted. So users are allowing control of their personal accounts by some untrusted source. An example of that would be, if I have a new photo on a subreddit, then put that as my phone wallpaper-- which could allow a potentially objectionable image to become your phone wallpaper without you authorizing it directly.

To summarize, we found that around half of IFTTT applets are potentially unsafe. And we say, potential, because there's still a lot of information about user-specific app settings and accounts that we're not taking into account in our analysis.

But these potentially unsafe applets can cause personal, physical, or cyber-related harms. We think that the security lapses we came up with are an interesting way to systematically reason about applets on IFTTT, or other similar services.

And we think that this work will become a foundation to create tools to help with awareness and decision making for users, which we think are clearly necessary. Because so many IFTTT applets do have this potential for quite harmful side effects. And you can read many more details in our paper. Thank you.

[APPLAUSE]

DAN SALSBURG: Thanks, Milijana. Now Gunes will talk about session replay scripts, and the research he's doing on that.

GUNES ACAR: Hello, everyone. So I'm going to present joint work with Steven and Arvind on session replay. So individuals watching over someone's shoulder, and taking [INAUDIBLE] up. This is like a common metaphor used to describe online tracking. But we know that it's not true. It's just machines like registering your visits, and [INAUDIBLE] crunching some data-- maybe profiling you.

So what I'm going to talk about today is, really make this real. So session replay scripts is a particular set of, class of analytics products. So what they do is, they record individual browsing sessions of you-- like your mouse movements, keypresses-- like, scrolling up and down. And they make it available as a recording, as a video to the web publishers, website owners.

So a company that we studied actually says, this is really like as if you are looking over their shoulders-- like visitors over a user's shoulder. So why do websites use session play scripts? On the left, you can see an example from a Yandex.Metrica, one of the scripts or products that we studied.

So you can see, for example, how the user's mouse moved over to the different buttons selections. So you can discover, basically, problems with your website. It's like analytics [INAUDIBLE] these products. But you can also detect the most interesting visitors. Such as the ones, for example, interested in buying products that you want to really sell. Or, you may detect why they had problems and didn't convert-- like, didn't really complete the check out.

So they really provide this useful feedback to the website owners, or the web publishers. But there's a privacy problem, associated with session replay scripts. So basically, these are recording or originating this session. Browsing sessions require a lot of data. First, it requires the scripts to grab this full page source.

That is, whatever you see on the page is collected and sent to this third-party server. In addition to mouse movements, and mouse clicks, and keypresses, session replay scripts have been known. Like, there are 10 years-old papers studying them. But they were mostly focused on watching mouse movements, or key presses. Whereas, we focus more on the collection of the page-- the page source.

So if you think a webpage you visit are just public documents that are the same for everyone-- that it's free of any private or personal information-- there is no problem maybe. The problem is not worse than any analytic scripts. But when you start providing data to the page-- like, input forms. For example, writing your name-- surname, social security number, credit card number, for example, for buying products-- then there's a problem.

The risk is that your personal information, or any sensitive information may be collected, along with the page source-- within the page source. This includes, like, email, credit cards, your passwords. So all these products, all the session-based scripts have some automated exclusion rules. Automated redaction, they call.

So we actually basically analyzed seven of them. And they all have-- like, based on either the type of the input fields. Like, for example, it's the password field, don't collect it. Or, based on the format of the data-- for example, if it looks like a social security number, don't collect it. So they have some measures against potentially collecting sensitive data.

Further, they allow the web publishers to exclude some of the input fields from the recordings. So if you're, for example, if you have input filtering our website-- to help people, for example, input a disease or the drug name. For example, you may just escalate that by just clicking and isolating it. It's very easy.

However, still, we found by just analyzing a bunch of websites we found session recordings that includes passwords, credit card data, student data, health data, and purchase details. So they were unintentionally collected by the session replay scripts, and sent to the third-party servers. So I'll just walk through some of the examples-- leaks or exfiltrations.

This is from a blog post that we published just two days ago-- Monday. So you may want to check it, because we have some other examples that we discovered in the meantime. Like, after we publish this talk-- specifically about passwords and leaks. So here what happens is, there's a third-party session replay script on this popular ads website.

So the session replay script is from FullStory, and collects all the data you enter into these forms. But they have some filters against, exclusion rules against password fields. But there is this show password feature, which just unhides or uncloaks the password, and displays it in clear text.

So whenever I click on this button, which is an eye icon so the stationary replay the exclusion filter just fails. So it doesn't recognize this field as a password field anymore, and just grabs the password-- as you see on the right. So we see that as you type your password, it just sent to the session replay company.

And normally, passwords are stored as hashed and salted. In that case, it is unintentional. They are just stored in the servers in plain text. So the reason for the unintentional collection is, when you have this Show Password feature, it just changed the type of the input field from a password to a text field. And when you convert it to text field, there is no way for the session script to really recognize that this is a password field.

So if you don't have this Show Password feature, are you free from any kind of problems? It's not. So we saw on this webpage from Capella University. So there is no Show Password feature. But this time, session replay script collects all the cookies and sends it to its server. For some reason it's configured to do so-- maybe to debug some errors. That only happens when you have some set of cookies.

And there is this other third-party analytic scripts, which stores your password in a session cookie. So one script stores your password in a cookie. And the other one just grabs it and sends it, which results in this password to be leaked to the session replay script server.

So the problem is not limited to password fields. For example, on the Bonobos website, they found that credit card details and other customer details-- such as your name-- is leaked to the third-party server. Again, this is just an unexpected markup, or composition of the page. That is different from what FullStory, or session filter expects.

Another example is, this time, related to health data. So Walgreens' website has FullStory, which is a session replay script again. So they basically manually redact all this potentially sensitive fields, such as doctor's last name, patient's name, or the content of the prescribed drug. But the name of the prescribed drug is not redacted-- possibly just a mistake-- and leaked to the third-party company. And on Gradescope, this time it is student details-- grades, names, emails, and commands are leaked to the session replay script.

So we find that these session replay scripts are pretty common. So almost 8,000 websites, and this is a lower bound, we think And just from manually reviewing a few thousand websites, you could uncover this-- like, very serious leaks, or exfiltrations. So we are just curious, how many more are out there?

So session replay scripts generally have precautions, but they are really fragile. And they can really depend on [INAUDIBLE] assumptions about the page. Redaction by the manual redaction is difficult and brittle. So the recordings contain really sensitive details, such as health data and student data.

So if you just think about if users are comfortable being watched, just read this Help page from companies explaining, is recording visitors legal? [INAUDIBLE] answers problem. And yeah, like all these first parties, or websites removed the session replay scripts since we published the study. But still, there are many more out there.

So you can find other similar research on the privacy problems of embedding third-party scripts on our No Boundaries series. And that's all. Thanks for listening.

[APPLAUSE]

DAN SALSBURG: Finally, Alan will present his research on how Facebook's advertising interface could be used to collect PI.

ALAN MISLOVE: Hey, good morning. My name is Alan Mislove. I'll be talking about joint work with my colleagues at Northeastern, at the [INAUDIBLE] Institute for Software Systems, as well as at UR Com. So I'm going to be talking today about Facebook and their advertising interface. But I'm actually going to start by talking about data brokers.

Now, you're probably familiar with these. But just in case you're not-- these are companies whose business model is to buy and sell data on people. They often buy data from sources such as-- voter records, criminal records, property records, and so forth-- aggregate and link that information, and then make it available in bulk to third parties. Often, clients include things like banks when deciding about loans, marketers, political parties, and so forth.

Now, what I'm going to argue is that online social networks and other similar services should be viewed as a form of 21st-century data brokers. The reason why is that all of these services are funded by advertising. They run powerful advertising platforms that they use to pay for the service.

Now, the way they work is they collect data-- not from public records, but instead from the online activities of their users on their sites, as well as on third-party sites. Again, they aggregate, link, infer information about that data. And then, make it available to advertisers, to choose which users receive their ads.

Now, the kind of data that you can infer from online activity is very different than the kind of data you get from offline activity. And so it's not surprising that today we're actually seeing the two of them partner. For example, Facebook now partners with Axiom, Datalogix, TransUnion, and Experian to allow advertisers to not only target Facebook-derived attributes, but also offline, more traditional data broker attributes.

And so the way that you do this targeting is crucial to understand our work. So I'm going to spend a couple of more slides explaining how Facebook's advertising interface works. So if you've ever placed a Facebook ad, you've probably seen this screen here. And so there's a number of different features, but there's three of them are important for this talk.

The first is highlighted in red. That's your attribute-based targeting. This is what you think of when you think of Facebook advertising. You can choose gender, age, interests, locations, and so forth, to choose which users are going to receive your ad.

Now, the second feature is a little one right over here highlighted in green-- where it's called the Potential Reach. Now, what Facebook tells you is, anytime you make a set of targeting selections, Facebook tells you the number of users who match your target. The idea being, if you're an advertiser, you want to know how big your audience is if you're about to buy some ads.

The third feature is this curious thing here at the top, highlighted in yellow. Now, this is called custom audiences. And the idea here is, this is a very different way to choose users. And in fact, instead of choosing users by choosing their attributes, instead you can choose the users directly. Meaning if you create a new custom audience, you get brought to this page-- where you can choose 15 different types of PII, that you can upload to Facebook and select which users precisely you want to target. This includes fields like email address, phone numbers, names, dates of birth, mobile advertiser's IDs, and so forth.

And so you literally tell Facebook what fields you have. You to upload a CSV containing the data you have. And then, Facebook will come back, match that against the Facebook database, and give you this thing called a Custom Audience that you can then advertise to-- containing only the users you specified.

Just like any other audience, Facebook gives you statistics on this. Meaning, how many users match? So of the records you upload, Facebook says, we found this many of those users. Now, I'm going to focus specifically on Facebook in this talk. But this feature is not unique to Facebook. Other services such as Google, Instagram, Pinterest, and LinkedIn all have these sorts of features-- where you can upload PII and target users.

Facebook's is simply the most mature, and allows you to target the largest number of PII attributes. So given that background, why are we interested in this? Well, when we discovered this feature, it raised a number of concerns. If you view these online social networks as 21st-century data brokers, essentially they have this large database of user attributes that they've built up for the purposes of advertising.

And when you allow advertisers to upload data and link against this database, that's a form of a database query. Meaning, I'm essentially allowing the advertisers to come to Facebook, upload data, and then we'll match it against the Facebook database.

And the thing is, on most of these services anybody can be an advertiser. All it takes is one click on Facebook, and suddenly you're an advertiser. So we're very concerned about whether this could be misused by malicious parties to inadvertently leak information-- because I'm being allowed to query this very extensive database.

So I'm going to show you that it is possible to do that. But to get there, I sort of need to walk you through a couple of different iterations, to understand how the interface allows it. So let's go back and look at this potential Reach Value. That's highlighted in yellow there.

Now, we did some studies of the interface. And what we found is that Facebook was rounding this number. It always ended in a 0. It was rounded to the years 10th. And ostensibly, the reason they probably wanted to do this was, to prevent you from learning information about one individual user-- because they're aggregating it. But I'm going to show you we can still do that.

So let's suppose you had an email address. And you want to ask, does this email address correspond to a Facebook user? So what you could do is, you take a list of records-- that's shown

here. Any records-- voter records, it doesn't matter. And you upload that to Facebook as a Custom Audience.

Now, Facebook's going to come back and say, OK, I matched 40 of those. The potential reach of that is 40. But remember, that's rounded. So now what you do is, you add one more record, and you upload that.

And Facebook comes back, and again says, 40. You add yet another record. Facebook says 40. And finally, you add yet another record. And Facebook comes back and says, 50. You've moved the rounding up. So what does that tell you?

Well, that second to last one is what we call a threshold audience. Literally, the true number of users who match is right before the rounding threshold. Meaning if I take my victim, and I add them to that threshold audience, I can look at the results. It's either going to be 40, or it's going to be 50.

But that leaks information. Because if it's 40, that tells you the user does not have a Facebook account. But if it tells me it's 50, they do-- because they pushed it over the rounding thresholds.

So this is the first way we can leak information, but it doesn't seem super powerful. But let's go a little bit deeper. So let's suppose you had a list of, say, a list of phone numbers. And you have your victim-- and you want to ask, is my victim's phone number in this the list that I have? So I'm going to call that my target list. Again, I've got my email address and a target list of phone numbers.

So, I play the same game. I start with my target list. I upload it to Facebook. And Facebook comes back and says, the potential reach is, say, 670. I add another record. It says 670. I add another record. And then it comes back and says, 680.

So again, the second to last one is my threshold list. The true number is right before the rounding threshold. So I take that-- I add my victim. And if I know my victim has a Facebook account, which I've determined from the previous step, I can then look at the result. It's either going to be 670, or it's going to be 680.

Well, what does that tell me? If it's 670, that tells me adding the victim did not increase the number of users in the list. Ergo, the victim was already in the list. Meaning, the phone number is in my target list. If it goes to 680, the opposite is true. I've added another user, and it moved up.

So now I can say, is my victim in a list that I've uploaded? Again, we're leaking data, but it doesn't quite seem super powerful yet. But let me show you how to you can abuse this.

Let's suppose you're smarter than your average bear. And you choose that target list to be, let's say, every phone in the United States starting with a 1. So 10 million phone numbers. And I ask, is my victim in that list?

Well, if the victim is not, I know their phone number doesn't start with a 1. If they are, I know it does. And I can play the same game with number 2, with different digits. And if I play this enough-- in fact, this is 100 lists-- I can infer the victim's entire phone number from doing this.

So this is an example of one attack. We have other attacks as well, that show that we can leak multiple pieces of user PII together. Let's say we can link email addresses and phone numbers to the same user. We can infer phone numbers, as you've just seen.

We can also deanonymize en masse all visitors to a website. Meaning if you put a tracking pixel on your website, you can then go to Facebook-- play these sorts of games, and get all of the phone numbers to everybody who visited your website from this interface.

It's important to note that throughout all of this we weren't actually placing any ads-- meaning, it didn't cost us any money. There's no actual interaction with the victim. Meaning that if you're this victim user, you can't detect that this happened. And you can't do anything to prevent it.

Now, of course this is a security vulnerability. We've reported this to Facebook. They put an initial mitigation in place. We're working with their security team to put a more permanent fix in place.

But the real sort of higher-level question is, we need to think carefully about these sorts of services. Because these are 21st-century data brokers, and because they have massive databases on people-- and they're using that to fund advertising services. And because these advertising services have numerous features-- more every day-- we need to think very carefully about what sorts of features might inadvertently be used to accidentally leak user information.

Now, this is all in an IEEE Security and Privacy paper that is coming out in May, but it's available on my website. We also have related work on Facebook looking at the explanations that it gives to users, and how Facebook advertising could be used by malicious advertisers to cause discrimination. If you're interested in all of those, they're all linked from the website there. Thank you very much.

[APPLAUSE]

DAN SALSBURG: Thank you, Alan. So we're going to start a discussion period now. If you have a question, you can fill out one of the question cards and hand it to one of the FTC staff who'll be walking through the aisles. Those were excellent presentations.

And as a consumer, they're kind of alarming presentations, too. And so I guess I just learned that if I open an email, it may result in my email address being sent off to who knows whom. That certain browser extensions are taking my browsing history and sharing them with people I don't know. That if I have an IFTTT recipe, it could result in my boss knowing that I really was at a party and not home sick.

And that third parties are seeing my mouse movements, and are also getting my browsing history. And that an advertising platform can, without my knowledge, be obtaining personal information about me.

So the question I have is, as a consumer is there anything I can do about this?

STEVE ENGLEHARDT: Yeah, I think the kind of go to answer in terms of web tracking is, you can run some type of tracking protection ad-blocker or something like that. Basically, if you prevent the request from ever going out, you can try to prevent some of that tracking.

DAN SALSBURG: These are generally extensions, though. So how would a consumer know that putting on such an extension isn't going to result in some other sort of privacy violation.

STEVE ENGLEHARDT: That's a good question. You've got to trust the developer.

MICHAEL WEISSBACHER: I guess there's a couple of highly-reputable extensions where we've found no issues. So uBlock Origin is an example of a privacy-enhancing extension where we found no issues with tracking. So yeah, uBlock Origin will remove a lot of these privacy issues, and [INAUDIBLE] and not leak.

DAN SALSBURG: So are there other things that a consumer can do? Alan?

ALAN MISLOVE: I'll say, on Facebook this is much more out of your control, in a sense, because it's data that they've collected on you-- oftentimes without your knowledge. I will say to Facebook's credit, that they do make some of this available. Some of the attributes, there is a Privacy Preferences page where you can get some of those attributes. And for some of the phone numbers, you can figure that out.

But what we found is that Facebook is not always as forthcoming as we would like. For example, you can never figure out the data broker attributes. They don't reveal those to you. And in fact, even in the ad explanations, you can right click on the ad and ask, why am I receiving this ad?

They will never tell you anything other than, it was an axiom data attribute. They won't tell you which one.

So, unfortunately, I think that's a case where consumers themselves can't actually do anything directly to do this. And instead, it may require either industry pressure or potentially some sort of government pressure, to get them to make it more transparent.

DAN SALSBURG: Milijana, is there anything that a consumer can do to find out whether or not an applet they're using has a secrecy or an integrity violation?

MILIJANA SURBATOVICH: There's nothing built in. To some extent, they have to think about it. Like, is there a potential harmful side effect? But these systems aren't designed in a way that's at the forefront of the user's mind. So, really, there needs to be more new things in place to adequately prepare users to be on guard against these violations.

DAN SALSBURG: And Gunes, anything to

GUNES ACAR: Yeah, like Steven said, extensions would be something I would suggest. But then, your question is kind of really [INAUDIBLE] I think like for the browser extension case, it should be like platform. Or maybe Facebook case as well, it should be platform providers-- like, platforms themselves having more stringent checks in place. Like Google for the Chrome extensions, or Firefox for the add-ons.

So maybe users just can ask for that-- like, having more protection against this kind of targeting, or disclosures, or this kind of malicious extensions. These are not the first time these extensions are found to be doing nasty things, I guess.

DAN SALSBURG: So, are each of your presentations really getting at flaws in privacy by design-- that the designers of these systems really weren't thinking about how their systems could be misused?

ALAN MISLOVE: I mean, I think in the context of Facebook there's really a tension between utility for advertisers and protecting privacy for users. Because the key thing that enables all of the sort of information leakage is that information about how many people were in an audience. And we show that we could abuse that in various ways.

And that's a trade off between how much advertisers understand about the platform, and the more that Facebook obfuscates the better protected the user privacy is. And so I think that sort of trade off is skewed a bit too much toward the advertiser than the user right now.

GUNES ACAR: So for the session recording scripts, it's not really they are trying to abuse something, but rather failing to protect-- failing to exclude the sensitive data from the recordings. We believe they don't really want to record this data themselves, because they are being liable for that. And that really brings some issues with that.

But it's just the model of all these wholesale data collection caused that. It's really very hard to protect against that.

STEVE ENGLEHARDT: Following up on Alan's point, I think in this email tracking, people-based marketing space there's this similar kind of tension between utility and privacy. So in order to do the types of tracking that these companies want to do, they need some way to in a decentralized way bring in a bunch of user data to a central location.

So I think the only way I could think to do that is, are these PII-derived identifiers? And when you read the marketing materials, as long as you hash it it's no longer PII-- it's safe to upload to us. And I'm trying to bring that into question.

MICHAEL WEISSBACHER: So browser extensions are just very, very powerful access to the browser. So usually, restricting anything privacy related will also harm extensions which want to do good things. So I don't think it's possible to easily restrict this technologically, just to block something within an extension.

So I think it's easier to be done by measurement, really. But it's impossible to just restrict this in a technical way, in editing extensions.

MILIJANA SURBATOVICH: And also for if this , then that, there's that large trade between convenience and privacy. Because as we try to show in the examples, there is an intended-use case for these applets that is quite useful and convenient. So it's not really a viable option for IFTTT to not allow users to make recipes that have these potential violations. Because that's half the recipes gone, right?

So I think there does need to be more informing users of potential violations. What's the worst case that could happen? It's not just the clear cut don't do this type of answer.

DAN SALSBURG: So you think it's some sort of pop-up in the setup process that would tell the users beware, make sure this is what you intend.

MILIJANA SURBATOVICH: Yeah, something along those lines.

DAN SALSBURG: If you have an understanding boss, that could be one of them.

[LAUGHTER]

So we have a question from the audience to you, Michael, about history-leaking extensions. Did you look for leaking of other sorts of data and other information-- for instance, the modification of pages that might be caused by the extensions.

MICHAEL WEISSBACHER: So I was looking exclusively at the browsing history. So pages that were visited, there's been other work that we're looking at whether passwords and other things are being leaked. But this work is only browsing history related.

STEVE ENGLEHARDT: Can I follow up with an anecdote on this?

DAN SALSBURG: Yes.

STEVE ENGLEHARDT: During my investigations of other things, I found in an example extension, which you log into. And then, in order to support the same kind of people-based marketing, the extension would inject pixels into the page that contained your email address hashed.

Very similar to what you see in emails, we saw an extension doing it. And I wonder how that plays with the kind of work that Gunes studied, where now you have PII in the page in an unexpected place. The site owner didn't expect it to be there, because the extension's adding it. So now it may even leak, aside from the companies it's intentionally being sent to, all of these third-party analytics as well.

DAN SALSBURG: We have another audience question, but directed at your research. And the question is, to do the type of attack that you describe to obtain consumer PI there are a lot of

steps. What kind of computing power is needed? Is this is something that is could happen in seconds? Or is this something that requires the resources of Northeastern and its computer center?

ALAN MISLOVE: So there were basically two phases. The first phase, we spent trying to figure out how the interface worked. And that took us a little while. Because we had to figure out, for example, they were rounding and so forth. But that was sort of all a one-time cost. So I think what the questioner is really asking is, at runtime, if you had you had wanted to attack, how long would it take?

So at the time we did the study-- Facebook has since changed their interface. But at the time, there is-- say, if I wanted to infer phone numbers. I have to upload all these lists of 10 million phone numbers each, right? That takes a couple of days, but that's a one-time cost.

So I upload those, and then I can use them from then on. To infer the phone number of an individual user, it takes about 15 or 20 minutes. And it can be entirely scripted. Meaning, their interface is just a bunch of AJAX calls. And so you can write a program that will just do this for you while you're making a coffee. And you can come back, and it spits out the user's phone number.

DAN SALSBURG: Let me ask a follow up on the Facebook issue from a consumer standpoint-- from my standpoint, my phone number's out there, and it's already being held by various data brokers. Am I harmed all that much more if somebody can just get the data for free?

ALAN MISLOVE: I would argue two things. The first is, that this is not unique to phone numbers. We didn't do it, but we have ways that you can use the same thing to figure out other PII about users, as long as you can sort of predict the PII. Things like email addresses-- I forget which one of the presenters were showing that email addresses can be predicted and so forth.

But you can use the same game to infer email addresses. But second, we are actually not worried about figuring out phone numbers en masse, but more for targeted attacks. For example, figuring out phone numbers of celebrities, or politicians, or so forth. That that's really the risk of-- there are sort of high risk people out there who might be victim to this, when the phone number is actually not public.

DAN SALSBURG: Steve has alluded to this, but can any of you think of a way to combine all of the different attacks that you've described to a nightmare scenario? Anyone want to take a stab at that one? We'll leave it for thought for another day, then.

The first question I asked was, what can a consumer do? What, more broadly, can the businesses do? I mean, a lot of these attacks-- it seems like there are businesses that probably don't want these attacks to take place. They don't want the data to be available in the way it's being made available. What steps should they be taking? Michael?

MICHAEL WEISSBACHER: So for these browser extensions, it would be nice if there would be some indicator that these extensions are leaking. Because for some, it might be intentional.

So there are extensions which save your browsing history somewhere else. So that obviously has to leak all the browsing history. But for some, it's not [INAUDIBLE] users might-- as long as they can make an informed decision, really, I think it's fine. The issue is only these extensions which are not upfront about these leaks.

DAN SALSBURG: So do you think that the browsers that offer the extension should be doing some sort of review, and informing consumers about what these extensions really do?

MICHAEL WEISSBACHER: Yes. So these stories have reviews. After uploading, an extension will be reviewed for several things, but privacy leaks are just not one of them-- or, not as far as I can tell.

DAN SALSBURG: Alan?

ALAN MISLOVE: And I'll say in the context of Facebook, I think the big thing that they can do is make it more transparent. Really, right now users have no idea how they're being targeted, why they're seeing certain ads. And Facebook does have some initiatives in this area. Like I mentioned before, they reveal some of your attributes, and they give you some sort of explanations. But it's not complete, and it's sort of hidden on the site.

There's one other feature that they have that they're deploying in Canada, where if you go to an advertiser, you can view all of the ads that particular advertiser is running. So that's interesting-- so you can actually understand sort of how people are using the Facebook advertising service. But they don't reveal the targeting parameters.

So it doesn't tell you who is receiving these ads, which is a really important piece. So I think making this transparent both to users, as well as to sort of third parties who can audit the services, would go a long way to addressing some of these concerns.

DAN SALSBURG: And, Gunes?

GUNES ACAR: Is there a reason? Like, why didn't they provide these parameters?

ALAN MISLOVE: I think it's privacy for advertisers.

[LAUGHTER]

DAN SALSBURG: Gunes, in the session replay script area, who are the players that should be doing something? And what is the something they should be doing?

GUNES ACAR: So this is kind of a debate between the session replay companies and the publishers. The session replay company says it's the website's responsibility to warn or inform the users. So maybe this is happening in the summer, down in the end-user license agreement that no one reads, or privacy statements.

So basically, I think it's the website's responsibility-- or maybe third party can have a an API or something-- to inform the users that their mouse movements, or keypresses are being monitored. And maybe the source code of the page-- like, pages collected. And I mean, this is another research, but I have a goofy design idea for that. Just like two eyes watching your mouse around.

[LAUGHTER]

I'll [INAUDIBLE] property.

STEVE ENGLEHARDT: Can I follow up on that? I think it's difficult, though, to say-- the problem is, even if websites attempt to fully audit their own site-- and say, OK, we know PI will be here, here, and here. A browser extension may change the site.

And if it does, there's no way a website's going to be able to predict all the changes that might happen from every browser extension. And so I think certainly that the session replay companies can move to a more-- for example, rather than collecting all input data, collect mass data.

So don't collect the actual text, but collect [INAUDIBLE] in place of it, would be a good way to prevent any accidental leaks. And there probably are other kind of architectural changes in that form.

DAN SALSBURG: Alan?

ALAN MISLOVE: I had a follow-up question on the session replay stuff. Because looking at mobile apps, I see a lot of similar things taking place there-- with these app analytics companies. Have you guys looked at that? I'm wondering if the same solution, or the problems that exist there may also exist in the mobile space.

GUNES ACAR: Yeah, totally. Like, we haven't looked into them. But we know that the same company is offering service about mobile [INAUDIBLE] also for mobile services. [INAUDIBLE] same thing, basically.

STEVE ENGLEHARDT: Yeah. And I'll say, that's a question we get often when we discuss the web results. And I think users have much less control in the mobile space-- including the mobile app space. In a web browser, you can install a blocking extension if you care. But in web apps, I think there are less options to do so.

DAN SALSBURG: Michael?

MICHAEL WEISSBACHER: Blocking by DNS is the only thing I could really think of. So within these apps, you can't install any ad blockers, things like that. But if you modify the DNS resolution, these requests to these trackers will go nowhere. But that's some of the work around [INAUDIBLE]

DAN SALSBURG: So that would be the website's responsibility?

MICHAEL WEISSBACHER: No, you have to do this on your devices.

DAN SALSBURG: So it's self help?

MICHAEL WEISSBACHER: Yes.

DAN SALSBURG: Milijana, are there things that can be done to help decrease the number of integrity and secrecy violations in the IFTTT applet space?

MILIJANA SURBATOVICH: Yeah, I think there are. Like previously mentioned, part of that is just the transparency. And, the notification if the user makes an unsafe recipe that they get that warning

DAN SALSBURG: Who would provide that? Would it be the manufacturer of the devices?

[INTERPOSING VOICES]

MILIJANA SURBATOVICH: Ideally, the platform would do that. Because how users use IFTTT is, they basically connect their service or device to it. And IFTTT asks for permissions to use that device. And then they make the rule on the IFTTT website itself. So it would have to provide that warning guarantee.

There's also engineering changes they can make, for instance. Like some of the labels, like the categories we provide, IFTTT could have some of that notion on their site. And then, users could tag their devices. Like, oh, I want this output from this device to be available to such an audience. So then it could immediately flag if there is some breach between audiences.

But that's like a lot of extra effort for IFTTT and the users. And the more effort you have, the less likely that users are going to use it. And they're just going to switch to a service that's simpler, easier, less hassle. So it's a balance.

DAN SALSBURG: So what makes PrivacyCon quite a bit different from the places that you've given presentations before is, that you have assembled before you policymakers, lawyers, staffers from the Hill-- people who are in a position to hear you and to cogitate on your information differently than a roomful of computer scientists.

So given that the audience is different than your usual audience, what would you recommend is the role for the FTC, or for other policymakers to address the issues that you've described in your research? And why don't we start at the end with Alan, and work our way towards Steve.

ALAN MISLOVE: Sure, I think the-- not to keep harping on this, but I think the key thing on Facebook is that we need to make the system more transparent. Facebook is in the news for a lot of other reasons-- but related to how their advertising system may be used or misused in various ways.

And so I think the biggest thing is to come at it from a view of, how would a third-party auditor be able to understand how the system is being used? And whether these features on the interface could be leaking information, but also could be used for other bad things, such as discrimination.

And what other problems might exist with the service? And right now, really we're relying on Facebook to give us certain views into that. But it's not complete enough to be able to fully audit.

GUNES ACAR: I think similarly for the session replay case, these companies could be nudged to be more transparent-- more upfront about what they collect, and the risks about this collection. So this could be user interfaces, or any kind of a [INAUDIBLE] that you can just go and see your data. But any kind of effort to bring more transparency and more control is, I think, welcome.

MILIJANA SURBATOVICH: You know, it's the same with IFTTT Because there is some, at least, efforts at transparency-- and like, Facebook, or Google, and whatnot. But IFTTT, or other similar services-- I'm not just ragging on IFTTT-- that connect the services, there's less transparency for them-- especially how they're connecting the services.

So, definitely policy could provide more incentive for IFTTT to be upfront about where data is going, and upfront about the risks-- as opposed to just pushing the innovation convenience aspects of it.

DAN SALSBURG: Michael, what should the policymakers here be thinking about doing?

MICHAEL WEISSBACHER: Well, I think it would be really nice if it will cause to incentivize these extension stores to scan for such leaks, and let the users make an informed decision. If they had the option to tell that an extension might be leaking, they maybe would not install it. Maybe they'd still want to install it, but it would be good if there would be transparency about this.

STEVE ENGLEHARDT: At the risk of repeating myself, I think really evaluate what's considered PII. I think, like I was trying to argue in the talk, hashed PII is still PII. It still can be used to identify the user. And I think that will change the legal structure, as to what can be shared-- as to whether that's considered PII or not.

DAN SALSBURG: Great. Well, this brings our session, our first panel to a close. Let me thank all the excellent presenters here, and tell you that we will have a break now until 11:20-- at which point, we'll promptly start with the second panel. Thank you again.

[APPLAUSE]

[MUSIC PLAYING]