PrivacyCon 2019 session 2

[MUSIC PLAYING]

JAMES THOMAS: Welcome back from lunch everybody. My name is James Thomas. And I'm an economist in the FTC's Bureau of Economics. My co-moderator is Jamie Hine, an attorney in the Division of Privacy and Identity Protection. And this session is on tracking and online advertising. You'll hear from five researchers who will each have 10 minutes to provide a summary of their work. And afterwards, we'll have a 20-minute discussion session.

So, while the questions will follow the presentations, please, start sending in your questions now-- either by writing on the cards and passing them to an usher or by tweeting us @FTC #privacycon19. So, we'll start by briefly introducing our presenters. The full biographies and funding disclosures are on our website. So, first, to Jamie's left is Catherine Han of UC Berkeley. To Catherine's left is Anupam Das of North Carolina State University.

To Anupam's left is Alessandro Acquisti of Carnegie Mellon University. To Alessandro's left is Cristobal Cheyre also of Carnegie Mellon. And, finally, we have Garrett Johnson of Boston University. So Catherine Han will start us off with her presentation-- "Do You Get What You Pay For? Comparing the Privacy Behaviors of Free Versus Paid Apps."

CATHERINE HAN: So, today in the advent of mobile applications, the mobile app ecosystem is largely defined by free applications. So, in order to get a better idea of what the consumer expectation is surrounding free applications and also their paid premium counterparts, we decided to conduct a survey. Our survey had 1,000 participants recruited from Prolific. And we wanted to start off by presenting a mock-up of two versions-- a free version of an application and its corresponding paid counterpart.

And in order to avoid any priming with regards to privacy or security, we started off with two open-ended questions without mentioning privacy and security at all. We first asked in which ways, if any, the participants expected the given apps to differ. And we followed up with a user preference question asking, which app they would be more likely to install and why?

So, based off our survey we discovered that without any priming whatsoever, only 1% of participants mentioned security and privacy as a possible difference that would be perceived between a free version of an app and its paid counterpart. Some even mentioned an explicit trade between security benefits and their willingness to do so in order to benefit from a free application. As far as user preference goes, we found that 20% of the participants were willing to or more likely to purchase the paid version of the application instead of the free one.

But 30% of those reasons were explicitly associated with ad removal in the paid version of the application, and only 6% mentioned security and privacy benefits as the explanation as for why they would choose to pay for the paid version. And some participants noted explicitly that they had an expectation surrounding the paid app having better security practices as opposed to the free version. So, that's without any priming whatsoever, but we wanted to dig a little deeper and figure out what consumers were expecting when it came to privacy and security explicitly.

So, what we found when prompting users to think about privacy and security, the story changed drastically. We found clearly that users were expecting worse privacy practices from the free versions of applications and, paralleling that, better privacy practices from the paid versions. For instance, in the free versions, users were much more likely to expect that the free versions would have a looser user data sharing practices, meaning that the free versions would be more likely to share user data with advertisers and law enforcement agencies.

In addition to that, users were more likely to expect the free version to also have poorer data retention practices, meaning that the free versions would be more likely to keep user data on apps servers for longer than necessary for the actual functionality of the application. Paralleling that-- for the paid versions of applications, users were more likely to expect paid versions to follow better security practices, which in this case meant encrypting data in transit using TLS. In addition, they also expected the paid versions of apps to be more compliant with laws and regulations, such as GDPR and COPPA.

And they also expected a higher level of transparency when it came to privacy policy disclosure from the paid versions of applications. And it's also extended into a higher level of granularity when it came to privacy controls when it came to the users interaction with the application. So, when it comes to this, the conventional media has been portraying this idea of if you're not paying for the product, then you are the product. And they're tying this implication of having a free application for this behind-the-scenes cost of the user privacy and user data.

So, some media experts have even gone so far as to associate this as a reason for why users should want to purchase applications, saying that users can now trade this upfront cost in a pay-for-privacy model, where they're paying for the product as a discrete application in exchange to avoid the data privacy violations that are occurring behind the scenes. So, in order to verify the validity of all these statements, we want to quantify how exactly the data collection practices of free apps actually differ, if at all, from their paid versions. And in order to quantify this, we looked across three different metrics.

First, we looked at the permissions declared. So, these are the permissions declared in the Android permissions system, which safeguards access to sensitive user resources, such as the user's camera, microphone, location information, et cetera. We also looked at the third party packages including the app, where third party packages could range from anything like graphics or utility libraries to advertising and analytics libraries, which we'll be focusing on today.

And, finally, we also looked at the different domains that received sensitive data, where domains are remote IP addresses that are receiving sensitive data in which sensitive data includes personally identifiable information, such as the Android advertising ID, user [INAUDIBLE], and coarse location information among other things. As far as our corpus goes, we constructed a corpus of 1,505 pairs of free and paid apps from the Google Play Store.

And our corpus currently spans over 1,000 different developers. In order to construct our corpus properly, we deployed an Amazon Mechanical Turk labelling task in order to find the correspondence between a free version of an app, and its intended paid counterpart. Because, currently, the Google Play Store does not contain any information or metadata linking the free version of an app to what its intended, if one exists at all, paid counterpart.

In order to conduct our analysis properly, we also wanted to ensure that we had side-by-side runs when it came to a dynamic analysis. What this meant is that we downloaded and installed the free version and the paid version of an application at the same time on a pair of identical Nexus 5X phones. And once we had those applications installed, we made sure to feed in the same random input stream of taps and swipes to each of those applications at the same time. This is the best effort of approach at controlling for any differences that we saw on behavior, but this doesn't necessarily control for UI differences between app versions.

Overall, here's what we've found. As far as permissions go, given that there was at least one permission declared by the free version of the app, we found that 79% of the paid versions declared the exact same, if not most of the same, permissions. One interesting thing to note is that 21% minority, where the paid version doesn't have any of the same permissions declared as the free version does.

And this hints at over permissioning that might be occurring within free applications requesting permissions that are definitely not necessary for the functionality of the app, given that the paid version definitely doesn't declare it. What we saw among third-party packages were that around 93% of paid versions, given that the free version had at least one third-party package bundled, also had some, if not all the same third-party packages. And as we mentioned earlier, third-party packages is a general category.

So, in order to get more insight into what these numbers really meant, we wanted to go ahead and categorize what third party package was included in which application. And for our study, we wanted to focus particularly on advertising libraries. So, in order to categorize the different libraries that we saw, we depended on pre-existing research using LibRadar.

So, based off of LibRadar's categorizations, we were able to identify 831 pairs of our corpus-- over half of our corpus. And we found that in this half, at least one of the free or paid versions had an advertising library. And maybe not so surprisingly, almost all of them-- almost all of the free versions had an ad library included. But somewhat more surprisingly is that almost half of the paid versions also contained ad libraries.

So, this is approaching a 50-50 shot of whether paying for an application is actually going to remove any ad libraries and if consumers are able to benefit by paying for the application at all. And this parallels what we saw with the network transmissions as well. What we see here is that given that there was a network transmission done by the free version of the app, we find that 44% of the time paying for an app is going to remove all of those transmissions and consumers are actually benefiting from this.

But that still doesn't explain why the other half of the time-- or around 56% of the time-- there's still data transmissions that are going on. And in 38% of the pairs, we found that all of the same exact data transmissions are still occurring. So, this brings to question whether paying for privacy is still a viable model.

So, overall, for the takeaways, we realized by doing our analysis that the measurable privacy benefits of paying for an application are tenuous at best. We saw, especially with the third-party advertising libraries and also in the network transmissions, that there seems to be a 50/50 murky decision of whether paying for an application is something that is actually going to benefit consumers when it comes to privacy. And because of that, consumers cannot reliably be expected to make an informed decision about purchasing an application because

none of this information is currently explicitly listed out on the Google Play Store. So, it seems that purchasing an app does not preclude you from still being the product.

[APPLAUSE]

JAMIE HINE: Thank you, Catherine. Next, we'll hear from from Anupam Das, who will be describing "The Web's Sixth Sense-- A Study of Scripts Assessing Smartphone Sensors."

ANUPAM DAS: Thank you. So this is a joint work with my collaborators-- Gunes Acar, who's at Princeton; Nikita Borisov, who's at UIUC; and Amogh Pradeep at Northeastern University. So, recently, smartphones have become the more dominant platforms for web browsing as this graph shows according to late 2016s. And mobiles have overtaken desktop in terms of the number-- in terms of the amount of web traffic that is generated by the different platforms. And this trend is expected to remain or even increase in the coming years.

And a result of this dominance-- and as a result of this dominance, we're seeing new mobile APIs appear. And some of these APIs enable developers and JavaScripts to access touch events, vibration events, and some of the on-board sensors that nowadays smartphones are equipped with. So, if you look at the right hand script, that is some very simple JavaScripts shows how easy it is to actually access some of the raw sensor data from your smartphones.

Now, in our study, we were only focusing on the raw sensor data. So, this included the orientation events, which basically tells you what the orientation of the phone in terms of the different axes, the motion sensors, which gives you data regarding the accelerometers and gyroscopes, and also the light sensors and the proximity sensors. And almost all of the major browser vendors support access to those APIs. But one caveat is that light and the proximity was only supported by Firefox at the time we were doing our study.

But one interesting thing is that in order to access these sensors, you do not really require any permission. So, if you look at the right hand side graphics, there is kind of showing you if you visit a particular website, you could actually tap into your sensors. And you can try it out in your smartphones if you wanted to.

We have the link in our demo scripts. So, you can see that as soon as visit the site, the site is able to capture the sensor data without invoking or asking for any kind of permissions to capture the sensor data. Now, you might be wondering, why would this be a problem if we enabled websites and scripts to access sensor data?

So, there's certain security and privacy implications in exposing your sensor data. So, for example, research has shown that by looking at your sensor data, you can actually figure it out or reconstruct the pins typed by a particular users. Accelerometers and gyroscopes have also been shown to act as a low frequency microphone, especially when you put them close to the audio source.

And using motion sensor, you can also surreptitiously infer the geolocation of a particular users, like the particular subway you're taking by looking at the motion-- looking at the curvature of the motion. And fingerprinting itself-- so, we ourself had our initial study in 2016 showing that just by tapping into the accelerometer and gyroscope data, you can actually fingerprint users.

A most recent study has shown that you can actually do this in a much larger scale. And the other one is biometrics. By looking at the sensor data, you can actually infer somebody's gait. So, in this paper, we kind of wanted to look into what kind of websites and what kind of scripts were actually tapping into or using the sensor data, for what purposes, and could we do anything to mitigate some of those risks.

So, we first had to go and kind of collect data. For this, we built our own-- we kind of built off of OpenWPM, which is an open source measurement platform. We instrumented over OpenWMP. But we did it for the mobile version, so that meant that we're emulating different kind of mobile environments but also meant that we were mimicking real sensor data.

So, we exposed sensor data and sensor data streams that kind of mimicked a smartphone being placed on a table. And we then was cautious enough to compare our fingerprints with real smartphone. And we saw that using the state of the art fingerprinting scripts, we were able to get an identical match.

So, once we've collected the data then we saw that, in general, out of the 1,000K websites that we visited, the different sensors were accessed in almost 3,700 websites, but the scripts were being originating from 600-odd domains. But if you look at the website themselves, you can see very popular websites, such as cnn.com, Reuters, or Wells Fargo. So, that means that large volume of user base is actually being exposed to this kind of information leak.

Then we kind of wanted to look into what third-party domains were actually accessing the sensor data. And then we concentrated on the motion and orientation because those were the-- we had larger number of scripts accessing the sensor data. So, the first one [INAUDIBLE] kind of an advertising company. And the second one-- [INAUDIBLE] is kind of a bot detection service. And DoubleVerify is kind of ad impression service.

And as you can see, a lot of those third-party scripts are loaded in websites that are ranked in the top 100. So, again, that means that a large volume of users are being exposed here. And we also found that some of the scripts were actually sending sensor data to the back end.

As I said, in our implementation, we actually mimicked the sensor data stream. So, we were actually-- we were in control of what sensor data stream we were generating. And we did this in a very crafty way, where we had some fixed parts, and we introduced some random parts. So, what that means is that we could actually automate this exfiltration process by searching for the particular pattern that we inserted. And we saw that a lot of the scripts were sending either the raw sensor data or encoding it into B264 format and then sending the data.

So, having known that scripts are accessing a lot of the sensor data, we next wanted to understand what was the typical use case for accessing the sensor data. So, for this, we kind of tried to cluster the scripts into different use cases. For this clustering, we looked at different low level features like what APIs they accessed and high level features like were the scripts potentially labelled as fingerprinters.

And then we kind of did an unsupervised clustering scheme using DBSCAN. And once we figured out the clusters, we then had to manually go through the scripts and individual clusters to figure out the use case. So, once we've done that, we found that a large number of the scripts can be categorized as tracking scripts because they were doing either fingerprinting or audience recognition or session replay.

The second most popular one was the fraud detection. So, they were distinguishing between bots. And also we found that some scripts were doing feature detection, gesture recognition, and some scripts were even used to generate random numbers.

But we then wanted to look into what fraction of the scripts that were accessing the sensor data were actually also doing fingerprinting. So, this was a very interesting kind of a table where we looked at individually the different scripts accessing the different sensors. And if you can see here that almost 63% of the scripts that are accessing motion sensor were also doing some form of fingerprinting, whether it's canvas fingerprinting or audio fingerprinting or battery fingerprinting. And these numbers are quite high across all different categories of sensors.

So, then we thought about, OK, what can we do here? So, the typical approach would be could we do some kind of a blacklisting or blocklisting. So, the blocklist would work, but it would only block the more prominent ones. But there's still the long tails. In our case, we found that the blocking rate was anywhere between 2% to 9%.

And also saw that some of the sites were actually loading the tracking scripts as a first party scripts, especially banking scripts. Those scripts were predominantly doing bot detections or fraud detections. The Feature Policy API is another approach, which could enable publishers to control certain APIs that can be accessed by third party scripts and their websites. But, again, this hasn't been adopted much.

And the other one-- the recommendation that WC3 makes is that why don't we block access to scripts from insecure or cross-origin iframes. For example, in our case, we found almost 63% of the scripts that were accessing sensor data was through cross-origin iframes. But it turned out that some of the most prominent, even the privacy geared browsers, such as Brave and Firefox were not following this kind of recommendation.

And there could be other ways to restrict it too. So, for example, by default we could lower resolution. You really don't need the fine grain sensor data to just figure out your orientation, for example. So we could also include some kind of indication in the browsers we currently have for speakers and cameras. In the privacy mode, we can, by default, just disable it. Because it is the private browsing mode we're using.

So these are some of the options that could potentially be explored. And, most recently, once we published our works in late February this year, I think Apple iOS came out with their iOS version 12.2 where by default, accelerometers and gyroscopes in the Safari is blocked.

So, if you tried to visit our demo pages and [INAUDIBLE] iPhone was using an iOS version greater than this then you probably didn't see any numbers there. And also Firefox, as I said, was the only browser which was giving access to the light sensors and the proximity sensors. As of May 2018, they've also kind of disabled that API. So, yes, with that, if you are interested in looking on into more of the findings that we found out or interested in using the framework that we built or the data or just want to know which websites are accessing what sensors, you can visit this website. And with that, I will thank you and end my talk.

JAMIE HINE: Thank very much, Anupam. Our next presenter is Alessandro Acquisti. He'll be presenting his piece on "Tracking Technologies and Publishers Revenues-- An Empirical Analysis."

ALESSANDRO ACQUISTI: Oh, thank you-- delighted to present this joint work in progress with Veronica Marotta and Vib Abhishek. I will start broad and then go narrow and then narrower yet, because we start from the broad research agenda that my research team and I have been trying to focus on for the past few years. To the extent that value and surplus is being generated by the collection and the analysis of consumer data, how is this surplus then allocated back to different stakeholders?

Often we hear that the data economy is producing some economic win-wins, where all the parties involved get a benefit. That may well be the case, but we want to understand better the extent to which different stakeholders are benefiting from this economy. And, therefore, we have a number of studies going on in parallel trying to piece together different angles on this economy.

For instance-- and here I'm going narrow-- if you consider online advertising, specifically online targeted advertising, there are at least two different ways to think about these in economic terms. One way is that there are consumers who visit sites, there are merchants who want to reach consumers. And the data intermediaries act as matchmakers between merchants on one side and consumers, publishers on the other.

Doing so, they reduce search costs on both sides of this market. And by reducing search costs for both sides, they create economic win-win. There is another, I would say, equally legitimate way of looking at things, which, again, starts from consumers and publishers and merchants. However, it realizes that consumers have a finite budget in attention.

They cannot look at all ads presented to them and buy all the products advertised to them. Publishers are under a conditionof extreme competition because due to the proliferation of different channels nowadays for which consumers can be exposed to ads-- not just websites but social media, text messaging, apps, and so forth. Merchants are also under a condition of high competition. Because if I was a producer of golf balls years ago, I may try to buy advertising on a golf magazine and compete against other golf balls or equipment producers.

Nowadays, I might try reach a visitor to The New York Times website because this visitor is interesting in golf. But, in fact, this visitor is not just interesting golf-- may be interested also in shoes, may be interested in the vacation to Cancun, may be interested in Italian sports car. And so many different merchants are competing for the visitor at the same time as they do. Now, in the middle, we have, of course, a very complex advertising ecosystem, which is however dominated by a few very large players.

So, If you have a two-sided platform with lots of competition on the side and some form of oligopoly at the center, you would expect more surplus to go towards the center. So, here are basically two frames and they're both legitimate. And my point is that we do have a little bit of data, but we still need more data to understand better to what extent each of these frames is correct.

So, going now narrower yet-- the specific study I presented today is just one piece of this much broader puzzle and focuses on publishers. The puzzle that we focus when we consider publishers is that we do know there are revenues in the advertising ecosystem have grown rapidly in the last 10 years, thanks to behavioral targeting. We also know from surveys, anecdotal data, and more surveys that not all publishers are doing so well under this system.

In fact, the revenues for publishers in some cases are stagnant; in some cases, according to them, even decreasing. And the famous case of The New York Times that in response to GDPR block behavior targeting, we found actually seeing a reduction in revenues is a telling case. What is up?

We try to understand it by looking at the relationship between behavioral targeting and publishers-- specifically, we do the following. And I want to be very, very precise in defining our research question, so that it is not either overblown or misunderstood. The research question we ask is, what is the increase in publisher revenue, after accounting for other factors, when the ads publishers are selling can or cannot be behaviorally targeted through cookies?

We focus on programmatic, open options, real-time bidding. And we exploit the fact that if the user cookie is available or not then audience based targeting would be possible or not. Of course, other forms of targeting may still be possible even when the cookie is not there-- for instance, contextual targeting.

There has been much work on the merchant side, the advertiser side on behavioral targeting. We do know that the behavior targeting increases click through rate and conversion rates. There has been a great work by Garrett on the ad exchanges. There is less work specifically focusing on publishers.

And on theoretical grounds, you could make somewhat opposite predictions about how publishers and revenues may change when behavior targeting is possible. One prediction is that when you can do behavioral targeting, the audience that merchants are addressing is more valuable because it’s more interested in a product, precisely due to targeting. This leads to higher bids by the merchants on their online ad exchange auctions. And these higher bids translate to higher revenues for the publishers.

There is an opposite story, which is the ability to micro-target audiences creates actually less competition about merchants-- between the merchants, I'm sorry, because it creates a smaller pool of subjects that any given merchant may be interested in. This may reduce bids. It may reduce, ultimately, downstream revenues for publishers.

The data we use comes from a large US conglomerate, which controls several websites. We have data on the ad features, where the ad was shown, characteristics of the visitors, such as the device type, et cetera, and, importantly, the revenue that the publisher received, as well as whether the cookie ID which makes this form of targeting possible or not was there. Now, the empirical approach is simple.

This is not an experiment, unfortunately. It's observational data. But we can distinguish revenues in the case that there is the cookie and there is no cookie. And this is importantly not a decision by the publisher. So, we don't have a problem of endogeneity there. It's the decision by the user.

And when we look at the raw means over the revenues that the publishers are getting, the raw means are indeed higher in the case when the cookies are present. You can see the CPM is $1.18 then when the cookie is there. It's $0.74 when it's not there. So it's an increase of about 55%, which is a large increase and in line with the conventional wisdom of higher revenues in presence of cookies.

However, the challenge here is that these are raw means, which do not account for other factors which may also influence the value of a visitor with cookies. You have to account for other forms of targeting which may take place, such as contextual targeting. We have to account for characteristics of the visitor, such as operating system, geolocation.

We have to account for user self-selection because the decision to have a cookie or not-- well, it's a user decision or the browser user decis-- or the browser set up which the user has chosen, this creates bias estimates if we do not account for self-selection. So, what do we do to account for that?

We use a technique called augmented inverse probability weighting. In essence, it's based on four steps. The first step is to estimate the probability model. The probability that any given user will have the cookie or not. Next, we estimate two different outcome models.

One for transactions with cookies. The other for transactions without cookies. And, finally, we compute the weighted means of the treatment specific predictive outcomes weighted with the probabilities that we obtain in step one. And then we, basically, take the difference. We compare the difference. And that is the result we're looking for.

This technique has nice feature called double robustness, which refers to the nice statistical properties of this model. And what we find when we use this technique is that, yes, there is indeed an increase, a statistical significant increase in revenues when the cookies are there. But it's smaller than what we expected originally. It is a 4%.

So, statistically significant-- economically significant, but still smaller than what we would have expected ex ante. Some limitations-- we do not claim that this is the overall aggregate value of behavioral targeting. We are just focusing on the very narrow aspect, what publishers are getting when tracking cookies are there. We're also analyzing the data from a single company.

Many, many different websites but one company, so it's not necessarily representative of the entire internet. We also observe publishers revenues-- net of all the fees the different intermediaries may be charging. We cannot comment on those fees because we do not track the rest of the funnel in this advertising ecosystem.

We also cannot capture the presence of more sophisticated or more invasive, perhaps, forms of tracking, such as device fingerprinting. Nevertheless, we believe that this result can help by adding one additional piece in this larger puzzle, which is the advertising economy. Thank you very much.

[APPLAUSE]

JAMIE HINE: Thank you, Alessandro. Next, we'll hear from Cristobal Cheyre. And he will be talking about his study on the impact of GDPR on the ad supported-content providers.

31:47

CRISTOBAL CHEYRE: So before I begin, I want to mention that this study is joint work with colleagues from the University of Paris Sud, Carnegie Mellon University, and the University of Minnesota. So, when we started this study, the motivation was that we wanted

to determine if the implementation of GDPR, which is something is good about protecting users' privacy, could lead to negative downstream economic effects by restricting the quantity and the quality of free online content on the internet.

The way this could happen is the following. There is still no consensus on how GDPR applies to the behavioral targeting advertising industry. But a common interpretation is that online data trackers must obtain user's explicit consent before tracking them. Users are typically tracked online in order to learn about their interest and serve them with behaviorally targeted ads.

From prior studies, we know that targeted ads are more profitable than non-targeted ads. And, thus, if GDPR restricts the number of users that consent to being tracked, it could make online advertising less profitable. We also know from prior studies that many websites derive most, if not all, of their income from advertising. Thus, it could be the case that the implementation of GDPR leads to online publishers making less money, which in turn could lead to some of them going out of business or the quantity and quality of their content being degraded due to cash constraint.

So, the motivation of the study is to see if this chain of events it could be in play or not. And to do this, we need to focus on both technical and economic variables. The technical variables are going to be related on how publishers implement the requirements of GDPR. And the economic variables are going to be related to the outcome of these publishers.

So, what we're doing is that we analyze a sample of over 6,000 websites. These websites are roughly equally distributed in the US and in the EU. And we started following them before GDPR was implemented. And we continue to track them until today.

From the technical side, what we're doing is that we visit each of these sites periodically-- roughly every two to three weeks. We use an instrumented browser. And we visit them from both EU locations and US locations. In each of the visits, we record what's going on-- for example, the number of first and third party cookies that a websites are placing in our client machines, the type of these cookies when we can determine it, the size of these cookies, the number of HTTP request, both first and third party, the size of these requests, in general how the website is recording information about the client.

From the economic side, what we're doing is that for each of these sites, we're collecting a series of variables related to the quantity and the quality of content. So, as a proxy of the quality of the content, we're looking at things like the reach of these websites or the number of page views that they're getting from each visitor. So, I mean, this should be roughly related to their quality in terms of that it reflects the impact that the content is having.

In terms of the quantity of content, what we're doing is we are looking over time how many new URLs are being added to domain. These should be roughly equivalent to the number of new articles being posted. And we're also computing some measurements on the characteristics of this new content. So, the first result I wanted to show you relates to the number of third-party cookies that websites are placing in our instrumented browsers. This looks only at European sites.

The blue line-- the line with triangle markers in case you're color blind-- shows the mean number of cookies that sites based in the EU are placing in our browsers when we visit them

from the US. The red line does the same thing. So, it's counting the mean number of cookies that EU sites are placing in our browser when we visit them from the EU.

You can see that when we started collecting this data, these two lines are pretty much a very close to each other. Approaching GDPR, the number of cookies starts going down. And after some time, it starts recovering.

What's interesting is that in the case of visits originating from the US, you can see that the number of cookies goes back to pre-GDPR levels. In the case of visits from the EU, it recovers, but it doesn't recover as much and stays below GDPR levels. What's more interesting is to look at what US-based websites are doing.

So, in this case, we observe the same pattern. Leading to GDPR, the number of cookies goes down. But after GDPR, the number of cookies placed on US-based machines, again, goes back to pre-GDPR level. But the number of cookies placed on European machines, it stays very low.

So, this is interesting in the sense that you can see how both EU-based sites and US-based sites are deciding what to do in terms of privacy of their visitors depending on where the visitor is located. In the case of US websites, they're being very careful in how they deal with European visitors. To dig deeper on what may be going on here, we look specifically at what some sites in our samples are doing.

We're focusing on news and media sites because these are websites that typically drive most or all of their revenues out of a advertising. Our sample has about 1,000 of these sites. 46% of them are based in the EU. 43% are based in the US. And we have 11% in other regions as a measure of control.

So of the US-based sites, we realize that roughly one in five of these sites are completely blocking the European Union. And then when we look why the characteristics of the sites that block or don't block the European Union, we realize that this seems to be a pretty rational decision. The sites that block the EU are sites that before GDPR were not getting that many visitors from the EU. So, these are sites that were getting 90% of the visitors from the US.

The sites that continue to allow EU visitors are those that were getting many visitors from outside the US. So, these sites were only getting 73% of their visitors from the US. So, as I said, it's quite interesting in the sense that the US websites seem to be being very careful in how they deal with European visitors. When they don't get that many visitors from out of the US, they seem to prefer to lose some of the visitors, to lose some of the associated revenues rather than having to deal with implementing the requirements of GDPR and exposing themselves to the liabilities of a potential data breach.

Now, from the economic side, the first result I wanted to show you is how reach has evolved over time. The graph here is a bit noisy. But doing an econometric analysis, we could determine that in the case of US websites, reach has declined a little bit after GDPR. I mean, this is hardly surprising as I've just shown you that about 20% of news and media sites are completely excluding visitors from one region. Whereas, in the case of the EU, it has remained mostly stable.

Now, looking at the engagement of visitors in these websites, it's hard to see in the graph because the effects are small. But using econometric analysis, we could determine that in the case of US websites, the number of page views per visitor has not changed much. Whereas, in the case of EU websites, there is a small decline in the number of page views per visitors.

We have not explored explicitly why this may be happening. But a plausible hypothesis is that, in the case of European sites, after GDPR, you see that there is all these privacy notices and consent mechanisms showing up. And it may be that visitors are turned off by these notices and prefer to leave the sites and go somewhere else rather than keep browsing that particular website.

Finally, in terms of new content being posted-- again, the figure is difficult to interpret. But when we do an econometric analysis, it seems that EU sites, instead of posting less content after GDPR, seem to be posting more content after GDPR relative to US websites. So before concluding, I want to mention a couple of limitations that we have to bear in mind when interpreting the patterns that I just showed you.

First, it has only been one year since the implementation of GDPR. The downstream economic effects that we're trying to measure may take a little longer to materialize than one year. Another problem is that there is still a lot of uncertainty on how GDPR should and will be implemented. We have not seen any enforcement action yet.

We're not seeing anyone getting fined because of the GDPR. Until that point comes, we're not really going to see what websites should be doing. Moreover, different countries have tried very different signals, with some countries you can saying that they are not really going to enforce anything for another year or so.

Also, our analysis is based on a subset of sites and a subset of metrics that may not be representative of the internet at whole and maybe our measures are not capturing the complexity of the technical changes being implemented because of GDPR. So, just to conclude, what we find in this study is that we are observing some technical changes being implemented after GDPR.

There are fewer third-party cookies and HTTP requests after GDPR. There is a recovery after some time, but still at least when browsing from the EU, we're seeing less cookies being placed on browsers. And this is particularly true for US-based websites. In terms of content, the quality of the-- and quantity has been-- we find these statistically significant results but the economic effects are pretty small.

If anything, it seems that the engagement at EU websites has decreased a little bit. And the way they have responded so far is by posting more content rather than less content as we expected initially. So, thank you, I'm looking forward for your questions.

[APPLAUSE]

JAMIE HINE: Thank you, Cristobal. Our final presentation is from Garrett Johnson. Garrett will be presenting his work on "Regulating Privacy Online-- The Early Impact of the GDPR on European Web Traffic and E-commerce outcomes."

GARRETT JOHNSON: Well, thank you very much. Today, I'm going to talk about some joint work with Sam Goldberg and Scott Shriver, where we are looking at the online impact of the GDPR. So the GDPR covers all of the EU, including the UK. And the GDPR is a very large, very complicated piece of regulation, but it is a generational shift in policy regulation.

And one of the key things that it does is that it increases the costs for firms of processing consumers' personal data. So, we think as a result of this, it's going to make it harder for firms to market to consumers using channels that use consumers' personal data. And it's also going to reduce the amount of data that firms collect about consumers.

Now, we're going to study this by using the May 25 of last year implementation deadline of the GDPR to do a before and after study and quantify its effects. Now, we're going to focus on its effects on web outcomes. And to get us there, we were able to partner with Adobe to get some really excellent data on 1,500 different analytics dashboards that tells you how many users are coming to websites, how much time they're spending on these websites, and, crucially, where they're coming from. So, if they're coming from the EU or elsewheres.

So, how does the GDPR impact the sort of data that Adobe would be able to look at? Well, it could affect it in two ways. First, it could change the total amount of web outcomes. And, second, it can change the amount of data that's recorded.

Now, the total web outcomes could be affected on the firm side by firms having a harder time marketing to consumers. It turns out that personalized channels like display ads and emails help drive traffic to these websites. And so, we could expect that that would put a drag on those total web outcomes. On the user side, I think Cristobal mentioned that there's this privacy saliency story, where if you're bombarded by privacy notices all the time, it might change your preferences for browsing online.

Now, turning to the percent recorded story. Firms may choose to record less data and no longer share data with Adobe in order to minimize their legal liability. But we are going to eliminate this by construction by just dropping all the cases where they stopped sharing data with Adobe from our data. There is, however, a consumer story here too based around consent. If consumers are not consenting to having their data shared with Adobe, then that data should not be showing up there.

So, as I said, the data we're going to use for this is quite extraordinary. It's 1,500 different analytics dashboards for large websites. And so, our data includes e-commerce companies, corporate websites, as well as content creators like news websites. And we have nice representation in terms of some of the largest websites online-- in particular, of the top 1,000 websites in the world, 128 of these are showing up in our data.

So, now, when we look at the average page views, before and after the GDPR, which is the blue line, we can see that page views appear to fall post-GDPR. But then the question is, are we capturing the effect of the GDPR or we just capturing the fact that Europeans are going on summer vacation? So, one way that we can try to resolve that problem is we can benchmark against what happened in 2017, which is the dotted line on the figure.

And you can see that these two lines track each other pretty nicely until the GDPR hits. And then afterwards, there seems to be this persistent decline in page views in 2018, which we

think is more attributable to the GDPR. So, using this strategy, we see a across the board reduction in recorded web outcomes. So page views fall 9.7%, visits fall 9.9%.

On our subset of e-commerce websites, we see a reduction in orders of 5.6% and a reduction in revenue of 8.3%. How big is 8.3%? Well, for the median firm in our data, this is worth $8,000 in revenue per week to that firm.

So, why is this happening? And to what extent is it total outcomes versus recorded outcomes? Well, one thing that we are continuing to work on is trying to understand based on where consumers are before they arrive on the website-- if they're touching an email, if they're touching a display ad or search-- if that could explain what's going on. But, unfortunately, we're still loading that data. So, I don't have results to share with you there.

We have, however, already ruled out the data minimization story that it's just that firms are not sharing this data with Adobe. If we were to include these companies that decided to turn off their data sharing, then the estimates we'd get on the last slide would be significantly higher. So, then maybe it's a consumer story. Maybe these privacy notifications or requirements of consent are changing the amount of consumers that are continuing to get their data collected.

But if that were to be the case, then we would expect to see that the type of consumer that is recorded in the data is probably different because we think that people that provide consent is different. And so, we would expect to see differences in the amount of page views that these users are viewing per visits and the amount of time they're spending on a website. And, in fact, we find that these commonly used metrics of user quality are pretty flat. So, there's really doesn't seem to be any movement there, suggesting that maybe this isn't coming from a user story.

So, the punchline then is we see this fall in recorded web outcomes on the order of 10%. It's hard to apportion how much of this is just an artifact of worse data and how much of this is an actual reduction. Certainly, if it is a reduction in total web outcomes, then that should be something that gives the regulator pause. Because this is bad for the health of websites-- e-commerce websites, news websites-- in the EU.

If it is a change in recorded data, then maybe that's of less concern. Maybe as the regulator, you want that to be larger. But certainly from the perspective of firms, firms are using this consumer data to improve the decisions that they make. And so, they're going to be able to make worse decisions as a result.

All right, so, I wanted to just take a couple slides to talk about a couple other papers we've been working on that speak to some of the discussion we're going to have. So, this other paper looks at the impact that the GDPR has had on third-party domains. So, it's essentially a very similar exercise to what Cristobal did of tracking all the third-party cookies and so on.

We're doing this across 28,000 top EU websites. And we find some really similar patterns. So, there's a drop in tracking or third-party domains at 14% the week after GDPR. Six months later, that's gone. The largest drop happens to be on the websites with the fewest EU users.

But where he takes this really interesting spin on the content that publishers are producing, we're instead focusing on the competition in the MarTech sectors that are represented by these third-party domains. And because the amount of this third-party domains is going down 14%-- like nobody is better off. But what we instead ask is, the pie is getting smaller-- is the pie becoming relatively more or less concentrated within the dominant firms?

And here we see for the top MarTech firms represented in our data, the categories like ads and web analytics and social media, we are actually seeing that there is a relative increase in concentration or relative decrease in competition. So, one cause of this is that when websites are choosing between duplicate vendors-- for instance, if they're choosing between Google Analytics and Adobe Analytics, they tend to choose the dominant firm.

So, in advertising, they're choosing DoubleClick 99% of the time, if they're dropping one. In social media, they're choosing Facebook 88% of the time. So, this is the sort of data patterns that would cause this increase in concentration. All right, finally, I wanted to talk a little bit about the value of a cookie.

We came to study this question in a slightly different way. And, actually, I just got the good news last night that this paper was actually accepted. So, it's going to be published in Marketing Science. So, what we did is we studied the industry opt-out mechanism, which allowed consumers who are concerned about online behavioral advertising to opt out of that. And the way that we study this is we were able to get data from a large ad exchange and look at these effects across tens of thousands of advertisers and tens of thousands of publishers.

So, what we find is that a very tiny minority of consumers exercised this opt-out choice mechanism. So, only 0.23% of consumers opt out. But those consumers that do opt out fetch much lower prices without this ability to target them. In fact, the price differences are 52% smaller prices, all else equal. Does this trickle down to the publishers?

Well, in our data, it appears so. Publishers are getting 40% less revenue from opt-out users. So, just by way of conclusion, I'll say that it's really exciting to be a PrivacyCon and to have a lot of energy and momentum around privacy these days and about privacy policy. But a recurring theme in the research that I do is that privacy isn't free. There are actual trade offs in the marketplace that result from privacy policies being implemented. And so, it's really important as we try to draft good privacy policy that we keep these trade-offs in mind. So, thank you.

JAMIE HINE: Thank you, Garret.

GARRETT JOHNSON: Thank you. I appreciate, sir.

JAMES THOMAS: All right, so we'll now turn to questions. If you are here with us today, you can please write your questions on the comment cards and pass them to us through the ushers. And if you're watching on line, please tweet us your questions @FTC using the hashtag PrivacyCon19.

So, first, this is a question primarily for Alessandro and Garrett, but also others, please feel free to weigh in. So, Alessandro's paper finds that cookies generate only modestly more revenue for publishers. But Garrett mentioned work just on that last slide that finds much larger effects of cookies on publisher revenue. So, what do we think can explain these

disparate findings? And if cookies only bring moderate benefits to publishers, who else in the online advertising ecosystem is receiving most of the benefit from cookies?

ALESSANDRO ACQUISTI: I'll start. Thank you. Well, I think there may be a number of different things going on and they wouldn't be mutually exclusive. If you look at the advertising ecosystem as a funnel with merchants on one side, the intermediaries, the ad exchanges, the platform in the middle, then the publishers at the very end, different studies have found slightly different values, depending on which part of the funnel you focus on. So, you go from studies that focus on merchants and find that merchants can pay up to 2.5 times more to target ads relative to not targeting ads. Down to what we are focusing on, which is not ad exchange data but these publishers data net of all the fees that publishers may be paying to the rest of the ecosystem.

The second factor could be that if you look at our raw means, they're actually in line with some of the literature. I quoted this raw mean difference of $1.18 versus-- I think it was $0.74, which is about more or less slightly above 55%-- less if you take a logarithmic transformation of the revenue. But still, it is a substantial difference.

But as I was explaining earlier, we have to dig deeper. And we have to control for other factors which may impact these raw mean differences, especially the self-selection by the users themselves. And that's where we arrive eventually to the 4%. And, by the way, that number itself is not a unique outlier in that other research-- for instance, there is this paper by a [INAUDIBLE], a very recent paper by the University of Washington.

They were not using website data, but they were using mobile apps data from a very large Asian network. And they found an increase in effectiveness due to behavioral targeting of 12% and then another 5% added when you account for contextual targeting. So, again, the numbers here are-- you can see all over the map, depending on the study, depending on their specific angle you focus on which suggests that these-- A, the results are context dependent; B, there is still much to understand in what I consider basically a black box economy.

I am referring to black box economy because sometimes even large players inside the economy realize only later on that they didn't know what was happening. Consider the scandals several months ago related to Facebook video analytics or the case or The Guardian suing a Rubicon for hidden fees. And, finally, a last possible point-- and I have to thank Garrett because we had the call some weeks ago and he suggested this.

As I mentioned in one of my last slides, one limitation of our data, which is great data, is that it comes from just one conglomerate, one media conglomerate-- many different websites but just one conglomerate. So, we cannot make claims about the internet as a whole. So, it's possible that we are not capturing what happens in the long tail of smaller players.

So, we cannot directly address that. I can tell you that after the call, we went back to our data to look for differences within our data between the larger websites and the smaller websites. And we found something that was surprising, meaning that actually the larger websites were the ones where the delta was actually larger in terms of revenues brought in by behaviorally targeted ads versus non-behaviorally targeted ads. So, again, I feel that there are different pieces of the puzzle that we are all trying to put on the table. And I do hope that these efforts contribute to eventually casting a light on the black box economy of online advertising.

GARRETT JOHNSON: So, I think we are in broad agreement about many of these points. So, one thing is that we're showing like a 52% difference on these ad exchanges for the inability to behaviorally track. But that isn't going to mean that it's going to be the same thing for all publishers. And you'd expect that a premium publisher, like the kind in Alessandro's data, should be able to fetch a higher price without needing to have this additional behavioral targeting information.

So, they should be less reliant on it. So, that makes total sense. And also I think we agree that there's much more of this on desktop than there is on mobile. And so our data is all mobile. And his, I think, more half and half. So, that could contribute to some differences.

One thing where we maybe disagree is just the role the intermediaries play in all this. So, yes, intermediaries take a share of the price that advertisers are paying for. But what we actually find our data is that the reduction when you lose this ability to behaviorally tracked someone is pretty split equally among these different intermediaries. Everybody falls by roughly the same percentage.

And if you think about the economics of this industry, it makes a lot of sense. Because, usually, the way this is working-- if you're an ad exchange, for instance, you're charging on the basis of per impression or you're charging a certain percentage of the advertising price. And that's basically what the other intermediaries are typically doing too is charging a percentage of the price. So that should avoid problems of like really skimming the cream causing some problems in the marketplace.

JAMIE HINE: So, if there are any questions from the audience, we can have those brought down. So, I'd like to actually sort of follow up but go in a slightly different direction as well. So, Garrett, both your research and Cristobal's research indicated that there were decreases in tracking immediately following GDPR but then some recoveries, which, in some instances, may have been almost to pre-GDPR levels after six months or so. And so, I want to follow up on what do we necessarily make of this recovery, also in the context of some comments, Cristobal, you made about not seeing enforcement or maybe there's some folks that are sort of waiting to determine whether to bring enforcement actions. So does this necessarily suggest that the negative effects of GDPR on revenue might be short lived?

GARRETT JOHNSON: So, I guess, I'll take a first stab at that. So, again, what we find is that there's a marked reduction in tracking like one week post GDPR and then it just disappears after six months. So, I don't think that this means that there's no impact of the GDPR. One thing that's important to realize that there's many GDPRs.

There is the GDPR as written. There is the GDPR as firms would like to believe what's in there. There is the GDPR right now, which is pre-enforcement. And one day we're going to see the GDPR with enforcement.

And so it would be very wrong to conclude based on fairly small moves in the data that we're looking at that there is no impact on the GDPR. There is just no impact of the GDPR without a lot of enforcement. But what is really interesting in our data is that these websites really are responding to incentives.

The websites that have the fewest EU users are actually the most aggressive about stopping this just because why risk a 4% fine on all of your revenues if a very tiny proportion of users

is coming from the EU? And we also see the guys that are moving the most in this data are the websites that have the most ads, the most content, the most words. These are the people that move down the most and also the ones that move back up the most.

Because I think that they really have a lot to lose here. And so, it makes sense that they would respond in fear to the possibility of being regulated against one week post but seeing no movement from the regulators would start to move back up again. So, yes, I think it's very premature to know exactly how it's going to shake out quite yet.

CRISTOBAL CHEYRE: So, mostly, I agree with everything that Garrett just said. I would add one additional effect. When talking with people from the industry on what may be going on and something that we have heard is that when GDPR came along, it brought a lot of attention to privacy policies, to cookies, and so forth. I mean, we all receive all the many updates of privacy policies. The same thing happened with cookies.

I mean, suddenly, all the technical departments of these firms were looking at cookies and started realizing that they had duplicate cookies. Cookies that they were not using anymore. Cookies of services that had expired and were not even effective. So, there was sort of a cleanup during that time GDPR was implemented.

But, now, after some time has passed, the same thing is happening again. I mean, cookies start getting accumulated because people don't usually a keep full track of everything-- all the third-party extension, all the things that they have installed on their websites. That could be one of the effects. And the other effect is the lack of enforcement.

There was a lot of terror when GDPR was going to be implemented on how strong enforcement was going to be. We have not seen any enforcement. So, it makes sense that some of these sites, especially the ones that are more affected, are starting to risk putting back all these things. I mean, we still don't really know and we need to continue following that. I mean, that's one of the motivations why we continue to run this thing and see what's going to happen once enforcement comes around.

ALESSANDRO ACQUISTI: Can I make a super quick comment tying it together Garrett and Cristobal's points. I agree with everything they said. It's interesting because all of us-- and there are so many GDPR empirical scholars in the room today. We're all trying to do this difference-in-difference, which is the typical approach that you would want to work in this case, but it's so hard because GDPR is actually a moving target. So, this makes it absolutely fascinating from the standpoint of a economic dynamics that's much harder for us to pinpoint precisely an effect. Because there is no single date where there is a off/on the switch.

JAMES THOMAS: Great, thank you so much. So, I think we're going to switch gears a little bit here with a the question from the audience that relates to Anupam's work. So the question is-- and everyone is free to weigh in on this as well-- but do you think that every form of browser and device fingerprinting should be banned? Or are there legitimate uses where these activities could be useful? So, for instance, fraud detection, audience recognition-- is there value to those capacities that might be acceptable if there is informed consent from users?

ANUPAM DAS: Yes, so, I guess when you talk about any scripts or application access in your sensors, the first thing doesn't really come to your mind these are being used for tracking you or doing audio recognition. So, obviously, that means there's a lack of understanding of

between the users and what's going on behind. So, that again comes back to the question of transparency.

So, obviously, if we think about having some legitimate use cases, obviously, the fraud detection seems to be one of the use case that you can argue for as being legitimate. Because you are doing this for security purposes. But, again, at this point, there's no kind of control or transparency to the end users to figure out or differentiate between what the purpose for accessing the sensor data. So, I think that has to be in place, even if we want to talk about this trade-off of benefits and the cost of releasing my private data.

And even in the context of broad detection, there should be some kind of control. So, for example, could browser developers maybe whitelist some of these vendors that are doing broad detections? Because in our study, we found that the two dominant broad detection companies were BC2 and Parametrics. And a lot of the financial websites like Bank of America, Wells Fargo were loading a lot of their scripts as first-party scripts.

So, it seems like there's few vendors that are doing this kind of broad detections. So, could the browser vendors may be whitelist and then just allow companies that are trying-- only companies that are within the whitelist to access the sensor data.

JAMIE HINE: So, Catherine, I have the next question for you. So, some of your results show that many apps do not change their data collection practices across free and paid versions. However, we might imagine that companies are not offering privacy as a premium feature because customers themselves are unwilling to pay that premium for privacy. So, I'd like you to maybe address what your results may say about this idea of a privacy paradox and the viability of pay-for-privacy consumer protection models more generally. We have a very similar question that came from the audience that basically asks, does this model work? And more so, should we even be putting a value on privacy at all?

CATHERINE HAN: Yes, I think based off of previous research, there actually has already been work done that shows that consumers are willing to pay a premium for privacy. But this is conditioned on the fact that the benefits of privacy are upfront and explicit for the consumer. And that's something that's missing from the mobile app ecosystem right now, even in the Google Play Store. If a version of an app has advertizements within the app and the paid version does not, all the Google Play Store denotes really is that one has advertisements and one doesn't. And there's no real connection there that's made with how that corresponds to user tracking. So. I feel like if there is some responsibility that lies on the platform provider to make explicit if there is a privacy benefit or not, I feel like there is some viability to the pay-for-privacy model. But I think the current issue is that there is no disclosure there, making that explicit for the consumer when the making the purchasing decision.

JAMIE HINE: Garrett, do you have any comment? You talked about privacy having a value cost in the context of GDPR.

GARRETT JOHNSON: I'll just say that when there's another great paper by Mike Kumar that looks at the similar sort of data set. He looks at how consumers respond to apps in the Android marketplace when they provide different permissions-- compares freed versus paid-- and does find that there is some sensitivity. So, it does change consumer demand if there's

more of these permissions. But the sort of imputed kind of willingness to pay that come out of that are pretty small.

Certainly, that's going back to the paper that we did on consumers that are trying to avoid tracking. We find a very small number that there's very few people that seem to be using this. And that just maybe because of usability issues. But then we looked at like usability figures for browser extensions like Privacy Badger that protect your privacy. And there's probably a lot of people in this room that use these sort of things. But I can tell you in the general population these numbers are really, really low. So, starting a privacy focused business I would not call a get-rich-quick scheme.

ALESSANDRO ACQUISTI: Can I tie together Catherine's answer to something Garrett mentioned at the very end of his presentation to the last part of the question, which I found very interesting. If I understand it correctly, the person was asking the very last part whether we should even try to put a value. I feel that Garrett was absolutely right when he pointed out that privacy creates trade offs. And we know this back from the days of the first scholars, who operated-- the first econ scholars who were operating in this area-- Chicago school economists like Posner and Stigler.

They were saying that privacy is redistributive. It creates economic winners and losers-- so, by the way, is the lack of privacy. There is no way out of that situation. What I feel we should always keep in mind is that as much as I find the economic analysis of privacy crucial-- and that's way I do work in that area. It would be hypocritical if I don't find them important. I do feel that they are only, once again, one piece of the puzzle.

There are things related to privacy, which are not quantifiable and we should not try to quantify them in economic terms because they are too important-- the mission of freedom and dignity and autonomy. So let's always try to do-- that's my view, my personal view-- our best to quantify what we can quantify. But always keep in mind that there are dimensions, which are not quantifiable. And we should not try to force a value on them.

JAMES THOMAS: Thank you, so another question building off of a Catherine's work. So, Catherine's survey results showed that users believe that paid apps have stricter data sharing practices. However, the analysis showed that that's not actually the case in many instances. So, the question is, should advertising a paid app as "ad free" but continuing to collect and share data be considered a deceptive practice?

CATHERINE HAN: Yes, I think part of what we saw in the survey is that though consumers weren't necessarily thinking about security and privacy, they were thinking about advertisements when they weren't prompted at all to think about that type of thing. But then, I guess, seeing how that has a gap between what we saw when we were specifically prompting for security and privacy thoughts. I think there is a bit of a disconnect between what consumers are considering advertising to be and how that relates to tracking, if at all. So, I do feel like there is some type of responsibility that needs to fall either upon the developer or on the platform provider in order to disclose what advertisement really means and what that can embody in terms of tracking and fingerprinting for the consumer.

JAMIE HINE: Actually, if we could broaden the question out a bit, and also, Anupam, you mentioned that a lot of the sensor data is not necessarily tied to a permission. So, if you could talk a little bit about your thoughts about the, I guess, lack of disclosure and maybe some

solutions about how to either improve that or whether there is a need for someone to sort of step in and better tie the sensor use to permissions or otherwise.

ANUPAM DAS: Yes, so I think if we talk about permissions-- I mean, the obvious solution is, as you see, is tying access to a permission. But we've already known from previous research that if we're giving too many options to the user, they're not really well equipped to make the right choices all the times. So, I think in that sense the default setting in the context of sensor data I think is as important. So, by default, what should we do?

And then, some of the recommendations that we make is that, by default, you really don't need higher resolution. Or in some cases, you don't even need access to data. In some of the real world use cases we saw were just gesture recognitions or responsive content display, which really means that you just need to figure out how you're holding the device or not. And those could be provided through other high level APIs, and you don't really need the low level APIs giving you the fine grained data.

So, I think the default settings could be one way of doing it. And that kind of loads off some of the decision making that the end users have to make. And sometimes they're not well equipped to make those decisions. So, that could be one of the options I think that we could go for. And I think that's what currently Safari has gone for by default.

And by doing so, we can also kind of push some of the tasks to the publishers or the developers, if they really require fine grained access-- for example, it could be an immersive online gaming. And so, you go to a website and you want to play game. In that case, we can ask the developers to explicitly ask the users for higher resolution data. And I think that makes sense in the current ecosystem because we're not seeing so much of those use cases right now. But if that was becoming the norm, then I think we have to fall back to a different policy.

JAMIE HINE: If I could just follow up and ask, so why do you think that that sensor use is not tied to a permission? Is this sensor use just such a new technology that it hasn't been tied to a permission? And if someone needs to sort of make that tie, is that the responsibility of the app stores or who should step in to do something like that?

ANUPAM DAS: Yes, so when we first saw that it doesn't require a permission, we actually were talking in the Chrome Developer Forum asking about why is this not under a permission model or anything. And I think one of the main reasons from the developers were that you don't want to invoke users each time they change the orientation of their phone, saying that, OK, you need to give me access to sensor data to figure out the orientation or something.

But our argument was that if that was the only use case-- if we do a use case analysis, does that mandate that you really required to fine grained access, right? At that point, we didn't see that. And that's why we made the recommendation that OK, by default, why don't you do that? And you can get orientation from any higher level API. You can just look at the width and height and figure out the orientation if you wanted to really.

JAMES THOMAS: So, we have a little more than five minutes left. So, I wanted to pivot back to GDPR for a bit, if that's OK. So Garrett mentioned this briefly in his presentation, but I wanted to talk about it more broadly as a panel. How has GDPR affected the competition among content providers and publishers in the online advertising ecosystem? And what can

we learn from the EU's experience with GDPR about constructing privacy regulations that promote competition?

GARRETT JOHNSON: So, I think it's hard. There's a theorized tension between privacy policy and competition policy. So just think of consent screens, right? It's going to be hard to consent to a very long list of companies. And consumers are probably going to be more likely to consent to a firm that they've heard of before and that's sort of inherently anti-competitive.

Now, again, what I found in our data is that everybody is worse off because-- well, the vendors are worse off anyways. I shouldn't say everybody. The vendors are worse off because a lot of them are getting cut from the marketplace post-GDPR. But they're not just being cut at random. It's definitely the case that these sites are favoring the dominant firms.

And so, these larger firms, even though they're losing an absolute share of the data, they're gaining a larger concentration within that smaller pie that's still left. So, that's, I think, a difficult trade-off because I don't think there's anything tremendously nefarious that's going on here. I think the sites are just choosing the vendors that probably have good market share because they offer a good product or because they think they're going to be more compliant with the GDPR.

But there's a quick thing I'd like to expand on is we also see some interesting competitive behaviors on behalf of the websites. So when we talk about the impact that the GDPR has on firms, one point that comes out a lot is that the larger firms have larger resources to comply with these laws. They have more engineers. They got more lawyers that they can throw at the problem than these small firms.

I think what that misses is that the small guys aren't in the radar of the GDPR regulators. The GDPR regulators don't go to sleep at night thinking I really want to get the number 1,052 website in Latvia. But Google and Facebook, it's not a matter of if. It's a matter of when. So this can create some interesting dynamics in terms of competition as well.

And, actually, that’s one thing that we've found pretty robustly in our data-- the largest decrease in these third-party domains is among the top ranked websites. Those are the guys that seem to be more afraid and seem to be taking more action. The smallest decrease in the short run is among the long tail websites. And, in fact, six months later, those are the ones that are driving this increase that brings us back to par six months later.

CRISTOBAL CHEYRE: One thing I wanted to add is that-- I completely agree with what Garrett just mentioned. I just want to add an additional dimension that we noticed. One of the things that we saw is there was a lot of concentration. When websites were reducing the number of cookies, they were-- So, if we observe a reduction in cookies, it doesn't necessarily mean that there is a reduction in tracking.

So, what we observe that there may be the same functionalities may be now being concentrated in a few cookies. For example, what we are observing is that a larger websites are reducing significantly the number of cookies, but those cookies are getting bigger. So, we still have to precisely determine what's going on.

But what we believe is happening is that the technical people at these websites are essentially concentrating what they had before spread into hundreds of cookies in just some few cookies.

I mean, it makes sense. It reduces the work that you have to do and it concentrates all the liability in just one component and not into multiple components. So, yes, there is definitely competitive implications that are going to come out of trust and out of convenience.

JAMES THOMAS: Great, well, thank you all so much for a great panel. Let's please give a round of applause to the presenters.

[APPLAUSE]

We're going to take a short break. Please, be back in the auditorium shortly before 3:30 for our final session. Thank you.

[MUSIC PLAYING]

ANDREA ARIAS: All right, if everyone could please take their seats. We're about to begin. If you've been here all day, I'm going to apologize because I'm going to introduce myself for the third time today. But for those of you just joining us, I'm Andy Arias. I'm an attorney in the Division of Privacy and Identity Protection at the FTC's Bureau of Consumer Protection.

My co-moderator is Lerone Banks. He is a technologist within the FTC's Division of Privacy and Identity Protection and our final session today is on "Vulnerabilities, Leaks, and Breach Notifications." You'll hear from four researchers. Their presentations will be approximately 15 minutes.

We'll conclude with about 20 minutes of discussions, where we'll identify some common themes and ask the presenters about their work and its implications. Again, we won't be asking questions until we get to the very end, but please feel free to start putting your questions down on comment cards. Just raise your hand and one of our colleagues will come by with a comment card for you to fill out. If you're watching us on the webcast, just go ahead and tweet us @FTC #PrivacyCon19.

So, let me introduce our presenters. Again, their bios can be found both on our website and there's some biographies outside in the front if you haven't seen it. So, take a look at them. So I'll just briefly introduce them now.

First, to my left is Sasha Romanosky of RAND Corporation. To Sasha's left is Elleen Pan of Northeastern University. To Elleen's left is Serge Egalman of the University of Berkeley and ICSI. And, finally, we have Yixin Zou of the University of Michigan. So, Sasha is going to start us off with his presentation on the creation of a model that can effectively identify vulnerabilities with a high risk of exploitation in the wild.

SASHA ROMANOSKY: It's a horrible title, isn't it? Trying to come up with a better one. Maybe we can crowdsource this effort. Actually, what this is is developing a better threat scoring system. But I'll get it into the details in a second.

This is a-- first of all, thank you for having me here. It's always great to be back. This is joint work with a number of folks-- Jay Jacobs and Wade Baker at Cyentia, who have really done some fantastic work. And a lot of this is really built on their work on analyzing data from many different sources including Kenna Security. Idris is also a co-author. But they really

have done some fantastic work in data analysis and risk analysis. And anyone who's interested in that space, I encourage you to follow them.

So, the story here starts from what I believe is a great failure of the information security field, of us as practitioners and of researchers, and our inability to answer very basic questions, very fundamental questions about security, about cybersecurity and risk. Are we more secure now than we were last year? What kinds of security controls should we buy? How much investment should we make in this world?

We're still not able to really do that. There are lots of different metrics that we can conjure up and we try and track that we think are correlated with these measures. But we're still not really able to do that. And because of this, firms continue to be breached over and over and over.

You've heard all of the stories. I don't need to tell you that. And I think this is important because it's not just a corporate issue of how to prevent these breaches and what can we do and what can't we do. It's not just a privacy issue. Because, at the end of the day, we all bear some of the harm from these breaches.

But it's also a national security issue, I would argue. It's a domestic security issue when we talk about critical infrastructure. And it's a national security issue when we talk about foreign threats that pose a risk to us as individuals, to our businesses, and to the critical infrastructure. And so part of the cause of this, I would argue, is vulnerability management. The ability for firms to figure out what they should protect, what they should patch, and how they should prioritize that patching.

I think firms are very good because of technologies and vulnerability scanners. They're very good at finding vulnerabilities. And, certainly, this recent wave of bug bounty and vulnerability disclosure programs have really helped that. It's really caused this excitement and this real interest in helping firms identify where the problems are but not in actually fixing them.

Now, sometimes, they can act as force multipliers to help firms address this. But, at the end of the day, those programs are just about the finding. And so it's that fixing part that we really want to try and help with.

And so, for many of the researchers on this panel, I'm sure today, the research becomes very important to us. We pour our hearts and souls into that. And for that reason, I think this research could have a very fundamental development, which is why I'm very excited about it. So, the ability to help firms better prioritize and better understand what to patch and how to organize that, I think, is a key issue.

The ways they go about that now are based on simple heuristics and severities. We want to understand if this vulnerability over here can really cause a full compromise of a system. OK, we should go after that. If this one over here just causes an intermittent denial of service, OK, we can leave that. So that's kind of how the prioritization goes about now.

But we would argue that what it doesn't include is information about the threat. Will this vulnerability actually be exploited in the wild or not? And that's what we're trying to develop here. And this has become even more important because of codified requirements by

organizations and by federal agencies to apply these basic heuristics, let's say, in their vulnerability management practices.

DHS recently issued a requirement for federal agencies to apply what's called the CVSS, the Common Vulnerability Scoring System, a way of ranking vulnerabilities, to their remediation of vulnerabilities in those agencies. In the credit card industry, the Payment Card Industry Data Security standard applies a similar kind of standard that all merchants that deal with credit card numbers need to show that they have removed vulnerabilities above a certain severity. So, it really becomes very important, I think, in order to figure out which vulnerabilities really are the important ones. Are they the high severity ones? Or is it another group of vulnerabilities?

And so, this is effectively what the firm's problem is. There is a large scale number of vulnerabilities that are known-- we're not dealing with 0-days here. But of those vulnerabilities that are known, only a small percentage are ever exploited. So, there is something on the order of 76,000 known vulnerabilities that have been identified and only 5% which are actually being exploited.

So, if you take, again, a common approach of using a vulnerability severity rating to fix those vulnerabilities that you think will be-- that score, say, an 8 out of 10 or higher, what you're doing is fixing a whole bunch of vulnerabilities, only a small subset will ever be exploited. So, it's a relatively simple problem, I think, to understand but identifying the key vulnerabilities that actually pose that greater risk is really the challenge.

And I think one of the reasons hasn't been until now is because, A, the data haven't been available. There haven't really been good sources of information about which vulnerabilities actually are exploited. There are many different organizations around that kind of collect little bits of information here, little bits of information there. But it really takes an organization, and people, and kind of the awareness to put all of that together to try and identify, again, which of these vulnerabilities will pose the greatest risk. And that's where we hope to make the contribution.

Now, another way that people may prioritize their vulnerabilities is based on published exploit. So, the story is here that either white hat hackers or researchers or whoever will find a vulnerability, find information about a vulnerability, and package that up in code, in malware, in an exploit and make it publicly known. So, there are some for-fee services and there's some open source services that provide this.

And this is part of the story of researchers sharing information about vulnerabilities, how they're exploited in order to help defend themselves. So, the story of-- so, what you might think is that vulnerabilities that are published publicly-- so exploits that are published publicly may pose a higher risk because then bad guys could take them and use them turn them into malware and lodged against companies to compromise the company.

So, what we might think of this is that firms might prioritize their efforts based on that. But, again, it suffers the same kind of problems that what you end up doing is fixing a whole bunch of vulnerabilities that you don't need to. So, there is this trade-off here. Essentially this is a classification problem. What you want to avoid are the type I, type II errors. You only want to fix those vulnerabilities that you know will be exploited and nothing else.

Now, there's a tension here and this is a longstanding debate that I'm not going to go into but I want to mention it of inference versus prediction. So, economists and those that use statistics will build their models, their empirical models, and they will include the variables that they think should have a good reason for being there because for they're interested in establishing causal inference-- that A caused B. Machine learning, AI, data science for large part turns that on its head and says, OK, we just want to fit the model.

We want to fit the data. We want apply whatever modeling techniques we can in order to achieve the best fit, the best identification, the best prediction. And these are really two fundamental camps. And that presents a tension for us because we want to do both.

We want to fit the data as best we can, but we want that to be very open and transparent. Machine learning kind of by construction is very black boxy. And that's a challenge for us. And so, what we're doing in this first effort is to provide the best fit that we can for the data. So, what we want to do is be able to say, OK, what is the best we can do at predicting these vulnerabilities that will actually be exploited in the wild? We'll have a separate effort, the effort that we're working on now, to open that up a little bit more to make it more transparent, to make it usable by everyone else.

Some of the issues with our data. There's what's called a class imbalance. And so what we have is a large collection of vulnerabilities, a large data set, only a small percentage of which are exploited as you saw before. That causes some issues for data modeling. This provides more information and there's lots more detail then you're interested.

We go through a lot of effort-- and again this is a lot of work done by Jay and Wade Cyentia and Kenna Security honestly to collect a lot of different data from many different data sources, information about the vulnerability, other characteristics about the vulnerability, links related to other descriptions about the vulnerability. For example, does this exploit a buffer overflow that could lead to a full compromise of a system? Is this related to a web application, a server application? All of those details, as well as the CVSS score, the scoring system-- its ranking-- and information about whether it's exploited or not.

This figure is relatively detailed. I don't expect you to read it. But what it represents and, again, that you can see it in full detail in the paper. What it represents are the results of the model.

So, what we've done-- the blue line is effectively our model. What we want to show is the best. And here we kind of demonstrate its performance overall. The axes are coverage and efficiency. And effectively you can think of it as the type I and the type II errors. What you want to achieve is the best coverage of all of the vulnerabilities that would be exploited but not patching those up will never be exploited.

And what we've plotted in the circles and you'll see in the labels below are different strategies. Different approaches for if you were to take this strategy of patching vulnerabilities, how would you perform? How many-- what would be the coverage of vulnerabilities-- and what would be the efficiency, the accuracy, for example, of that strategy?

Now, it's nice that our model performs the best overall. And we think that's very useful. The size of the circle represents the number of vulnerabilities that you would have to patch in

order to achieve that strategy. And so, of course, what you want is to be on the highest level-- more to the right and more to the top with a very small circle.

And so, from our initial analysis, I think we've achieved pretty good results of identifying strategies that perform well. Now, this is somewhat to be expected if you throw everything into the pot and let it churn train on the data and test on the data, you would expect to achieve some good results, and we do. But it's nice to see exactly how that performs.

So as I mentioned, what we have here at the end of the day is what I think of as a very fundamental step, a very important step in improving and evolving our understanding of risk management and, again, from a national security perspective, from a privacy perspective, and a corporate business perspective. This is part of the evolution of our understanding, first of all, that firms as an industry, we are not very good, as I mentioned, at assessing risk and describing this risk and understanding, again, how well we are doing relative to next year. Part of that is wrapped up in this vulnerability management strategy.

Part of it is wrapped up in understanding what really is the severity of a particular vulnerability. So, we're based on-- as I described, we're based onright now, we're using very simple strategies of severity. But I think what we're trying to do here is really move it to the next level and really try to improve all of our practices, which, hopefully, should improve everyone's understanding and increase the security posture for all companies.

So, stay tuned for more information. What we want to do at the end of the day is make this, as I mentioned, a threat scoring system that is usable for everyone, that is not just a proprietary black box. And two of the authors will be presenting this at Black Hat later this year. Thanks very much.

[APPLAUSE]

LERONE BANKS: Thank you very much, Sasha. Next, we'll have Elleen Pan, who will talk a little bit about her team's observations of Android's applications and their access to audio and video information.

ELLEEN PAN: Hi, today, I'll be presenting our study characterizing audio and video-- oops, sorry-- audio and video exfiltration from Android applications. Multimedia sensors on our phones have given rise to this persistent rumor that our mobile apps are constantly watching and listening to us. There are examples of mobile apps that have done this to backup these types of claims.

For example, SilverPush was using ultrasonic beacons to do cross-device linking. Facebook has filed patents to recognize user emotions as a scroll through their news feed. There was a soccer app that was using the device microphone to listen for unlicensed broadcasting then using the location data of that device to figure out where they were coming from. And there have also been examples of photos being taken surreptitiously by shrinking the preview window down to a one by one pixel.

Companies have a lot of incentives to understand and potentially control their users better, but media surveillance, thus far, has just been anecdotal. So, some of the goals of our study are to identify and measure media exfiltration at scale, meaning we use a large number of apps and we also broadly cover the app stores. We also focus on exfiltration over the network

as opposed to privacy risks caused by app access to the hardware itself, like location tracking or device fingerprinting.

And we also determined whether that exfiltration should be considered a leak-- that is, undisclosed or unexpected. We're also interested in how apps use sensors-- so, the permissions that are requested, APIs that are called, and whether those APIs are called by first or third parties. Third parties being things such as ad libraries, analytics libraries, et cetera.

So, for the purposes of our study, we define a media leak as one that is suspicious or unexpected. And to do this, we ask four questions. Users don't expect media shared outside the primary purposes of what the app does, thus presenting a privacy risk. If undisclosed to the user not only is it unexpected, but it might also violate privacy law.

Many recent pieces of legislation, such as GDPR and CCPA, require detailed disclosures of PII that is collected and used. Media shared outside the normal functions of what similar apps do is a good indication that it might be suspicious. And, lastly, if unencrypted over the internet, eavesdroppers can easily pick up on media that is shared over the network. So, if the answer to any of the above questions is a no, we consider it a leak.

Given the motivation and the threat of this issue, we developed a methodology to filter apps, collect traffic, and detect media leaks according to this pipeline. Our first step is app selection. In our study, we only looked at Android apps since Android is a platform where apps are the most amenable to code analysis and automated interaction. It was also not feasible to test every single app since there are more than two million apps in the Google Play Store alone.

Instead, we chose a subset of popular, new, and random apps from the Google Play Store and three popular third-party app stores. And we filtered them based on whether they called camera or audio permissions. And this totaled 17,260 apps. We use a large number of apps and samples from different stores in order to achieve our experiments at scale and to also broadly cover the app stores, since the official Google Play Store is more tightly controlled than third-party stores.

Our next step is static analysis. Static analysis helps us understand the privacy implications from the app code. Our static analysis consisted of permission analysis and API reference analysis. Permission analysis consisted of the camera and record audio permissions. And we look in each app's Android manifest file to see if these permissions are requested. For media API reference analysis, we used Android standard camera and audio API calls outlined in their SDK.

And for screen capturing-- since there isn't an SDK outlined way of doing this, we just use the most straightforward code that would programmatically capture screenshots. We then decompiled the apps and searched for these method calls. And for third-party media API references and third-party libraries, we rely on the third party package names and again search for these APIs calls.

Static analysis gives us a large set of apps that are capable of recording audio, images, and video, but they don't actually tell us which ones leak media when they are used. To address this, we used dynamic analysis, which consists of actually running the apps to see if they leak

media. So, for dynamic analysis, we use a testbed consisting of 10 Android phones, each performing automated random interaction with each of the apps.

And we use real Android phones rather than an emulator to avoid cases, where apps are programmed to act differently when emulated. We then recorded the network traffic using a man-in-the-middle proxy. And we extracted media from the network traffic using file magic numbers. And we validated our methodology and results by using a test app that we developed and known apps that we expect to send media over the network.

And we also verified detected media by manually interacting with the apps and replicating the leaks. This gives us our actual media leaks. Our research methods are rigorous, but they still might yield false negatives, since apps might leak media in ways that are static and dynamic analysis did not detect. Thus, our findings are an underestimate of the prevalence of media leaks.

However, our results do cover popular apps, so they speak to the commonly used ones. And we also ensure that we don't have false positives in our results, since we manually validated each case. So, for our results, we find 21 cases of detected media, 12 of which we consider leaks under our previous criteria of being unexpected or unencrypted. And we find that 9 of these are shared with third parties.

The small number of leaks is good news. We find that media links are quite rare-- only 12 cases out of 17,000 tested apps. However, they are not 0, meaning that such auditing is extremely important. Our first case study are photography apps.

We found that a slew of apps were performing server-side photo editing, meaning that apps were sent to the servers to get processed without any notifications to the user. In all cases, the app had no other functionality that required an internet connection, such as social media sharing or downloading new filters. And in five of the apps, the privacy policy vaguely disclosed some kind of personal data collection but didn't make a specific mention of collecting photos at all. And one app didn't mention collecting personal data at all.

Our second case study is a type of privacy or media exfiltration that we didn't anticipate at the beginning of our study and is also potentially an incredibly invasive privacy risk. And this is screen recording. We found that an app called goPuff was sending a screen recording of user interaction where PII was exposed-- in this case, a zip code. And this was leaked to a third-party domain belonging to Appsee.

Appsee is an apps analytics platform that touts screen recording as a feature, but places responsibility on developers for hiding sensitive screens. However, we found that few apps actually use the API method of doing so. And, although, there's a server side way that exists, it's unknown to us how many apps use it. We responsibly disclose this type of behavior to all applicable parties.

goPuff pulled Appsee from their Android and iOS builds and updated their privacy policy. Google reviewed the two parties and gave us this statement, and they also removed additional apps beyond our findings that violated their policies. And after some back and forth regarding the privacy implications of this type of behavior, we were ultimately met with no response from Appsee.

Our work was covered in the press, and it was largely motivated by this question-- is your phone spying on _ on one hand, many were relieved to find that we did not see any cases of audio being transmitted. However, the screen recording behavior was alarming enough that many still constituted it as spying. Given this type of alarm, access to the screen should be protected by the operating system. Or users should at least be notified and be able to opt out by not installing the app.

Main app and third-party permission should also be separated, since it is unlikely that third parties will require all the permissions that are provisioned to the main app. And there's also a need for independent automated testing to continuously audit apps. In conclusion, we find 12 cases of unexpected or unencrypted media, 9 of which are shared with third parties. We also find an alarming case of screen recording video that is sent to a third-party library, including sensitive input fields with no permissions or notifications to the user.

And this type of behavior is alarming because it's akin to having a third party looking over your shoulder as you interact with an app and memorizing stuff that you type in but never send, including credit card numbers, passwords, and unsent messages. And we focus on Android in this study, but there's more work that needs to be done for iOS. Although, the screen recording behavior that we found was also found in major iOS apps earlier this year. For more information on our study, you can visit our website. Thank you.

[APPLAUSE]

ANDREA ARIAS: Thank you, Elleen. We'll now hear from Serge Egalman. He has uncovered a number of side and covert channels in active use by hundreds of popular apps and third-party SDKs in the Android ecosystem. So we'll definitely hear from you.

SERGE EGALMAN: OK, thank you. And, yes, so I guess I'll just get started. So, we've heard a lot earlier today about Android permissions. So the permission system is shown to users whenever apps try to access sensitive information on the device. So, this is basically a way of supporting notice and choice, so app developers can provide notice using these permission dialogs and users, ostensibly, can read these and make decisions about their privacy.

So, this governs access to all sorts of data on the device, including sensor data, such as the GPS sensor or the camera, as well as various things like the file system, where the photo library is stored, as well as persistent identifiers that could be used for long term tracking. So, the question is, does this work in practice? We've been building a system which basically gives us a pretty unique end-to-end view of what apps do with sensitive user data.

So, this started initially as a project to support several user studies where we added instrumentation into Android-- basically, just to look at how often the different permissions were access by apps. And so, we rolled our own version of Android that has this instrumentation. So, every time an app tries to access sensitive resources, we can log that.

We then bundled this with some custom network monitoring code that then allows us to see the traffic. So, basically, we can see when sensitive data is accessed and then to whom it's transmitted. And we've bundled this all together to basically build a pipeline to automatically examine the privacy behaviors of apps. So, we've been building up an app repository that has about 100,000 unique apps.

We've been doing this for about three years now. I think we have a total of about 300,000 unique versions of all of the apps. This then gets fed to the test bed as we encounter new versions of apps. We run it on the phones with our instrumentation. And then we simulate user interactions by essentially generating random UI events on the screen.

We then take all logs from the instrumentation, and we have a database that allows us to query what a particular app did. We have a website you can go to. If you go to search.appcensus.io, you can search for the privacy behaviors of various free apps. But this is currently in a state of flux. As Justin Brookman alluded to earlier, we've spun part of this off as a startup. And so, right now, we've been focused on the back end to make it more scalable. So, the usability of the website is going to be updated soon.

Anyway, previously, we did some work looking at privacy compliance, but now we've shifted to look at outright deceptive practices. So, whenever you have a security mechanism, the security mechanism is, obviously, only as good as it prevents users from getting around that security mechanism. And, obviously, this applies in the physical world as well as in various technologies as well.

So, the two things that we are looking at were covert channels and side channels. So, a covert channel is basically-- imagine you have a security mechanism that protects access to sensitive device resources, such as location data or the microphone. App 1 might be allowed access to those resources because the user is granted the permission. App 2 might have been denied access because the user didn't want to grant permission.

App 1 could communicate with app 2 share with it the information that app 2 is otherwise forbidden from accessing. That's known as a covert channel. And side channels, on the other hand-- basically, there's the security mechanism. But if there are ways of driving around that security mechanism, that's known as a side channel.

So, using our infrastructure, we have the app database. We have the results. We can then do queries to try and look for the presence of various side channels and covert channels by, for instance, querying the number of apps that have been transmitting various types of PII and then looking at the number of those that didn't actually have permission to access that PII.

So to give an example of how that looks, imagine we have a set of apps that have not been granted access to the location permission as well as a set of apps that are transmitting location data. One would expect that the intersection of these two sets would be 0. That is, in fact, not the case, and it was that observation that had us looking around to try and figure out how it is that apps are accessing this data.

So, what we do is we compute how many apps are accessing data that they don't have access to. And then for each app that appears to be cheating the permission system, we then reverse engineer it. So, all the other stuff up until this point is automated. This, however, is a little bit of a manual process because it involves decompiling the apps and reading through assembly code to try and figure out what it is that they're doing exactly.

However, once we're able to do that and we identify the mechanism that it's using to get around the permission system, we can then create a fingerprint of that and then quickly scan the entire app corpus to see in how many other apps that same code appears. And so, it's sort of a semi-automated process.

So, into the details. So, most of the ways that apps seem to be getting around this relies on one fundamental issue, which is that while the Android APIs are protected by the permission system, the file system often is not. And so, there are apps that can be denied access to the data, but then they find it in various places on the file system, which they have full access to.

One of the ways that we observed it is in the proc file system, which for those who are familiar with Linux, basically, proc is a virtual file system that creates a directory tree and there are files there that reflect system state, such as various hardware information as well as networking information. So, one of the types of data that's protected by the Android permission system is location data.

And that includes not just GPS location data but also information about the Wi-Fi router, known as the BSSID is the Mac address of the upstream router. And we found-- I mean, it's pretty well known now that that's actually a pretty good surrogate for location data. And, in fact, the FTC has gone after companies who have been collecting BSSIDs in lieu of location data and without user permission.

This, in fact, is located on the proc file system. And so, there are many apps that we observed which try to access the data the right way through the Android API and then, failing that, try and pull it off the file system. So here's an example of one particular SDK, which monetizes location data.

At the top, it's trying to see whether the app has the access Wi-Fi state permission, which is what one would need to collect the BSSID of the router. And then, if it fails that, it jumps to this other function, which then opens up the proc file system and just read it there. So, what this means is there are situations where the user might have been prompted explicitly to grant location data to the app, they decline. And then the app reads that that data off the file system instead, which seems like that is a pretty deceptive practice.

And despite that, it's fairly common. So, we found lots of different SDK that were exploiting this particular vulnerability. The number of apps that are using these SDKs are-- I mean, there are hundreds of apps that are exploiting this. But the user base for those apps are in the billions. And these SDKs are developed all over the world.

So, another vulnerability that we observed is-- another way of collecting location data is just by connecting to the router directly. So, we found apps that would-- after being denied access to location data through the Android API, they'd connect by using the Universal Plug and Play protocol, which is normally used for configuring a wireless router. But if you connect using UPnP, the device will yield information about itself. So, these apps were actually making direct connections to the upstream router and then just reading the data off of the router using this protocol.

Another one we found was access to the IMEI, which is a hardware-based identifier, which, unlike the ad ID, can't be reset. It's in hardware. So, this is protected by the phone state and identity permission in Android. We found at least two SDKs that would be in some apps that had been granted this permission. And if they were in an app that had been granted this permission, they would then write the IMEI to the file system in a publicly routable location so that other apps containing that same SDK if they were then denied the permission, they would then go to check for this file to see if it was written by another app and then grab the IMEI that way.

And, again, this corresponds to about a billion installs of the various apps that are exploiting this technique. Another one-- Mac address, another hardware based persistent identifier, which now in current versions of Android, this is totally off limits. There's no Android API that allows access to this data. Because, again, it's a hardware based identifier that can't easily be reset.

We found Unity is exploiting C++ libraries on the device to collect the Mac address. And we observed this in over 12,000 apps that we're using various Unity SDKs. Pictures on the device-- so, photos contain metadata. Sometimes the metadata includes location coordinates. We found that the Shutterfly app, which had been denied the location permission, was opening up the photo library on the device, reading the exif metadata and then sending GPS coordinates to its home servers.

So, the conclusion is that the Android permission system is designed to prevent access to this personal data or, at least, allow users to regulate it. But when the same data appears elsewhere on the device and it's completely unprotected, apps will and do get around the permission system. So, we reported this to Google back in, I believe, September of last year. They awarded us a bug bounty.

And they claim that this is going to be fixed in Android Q, which supposedly is going to be released sometime in the next year or so, which is good. But when it comes to security vulnerabilities, they have over-the-air hot fixes that address many of the security vulnerabilities. At the same time, Google is publicly claiming that privacy should not be a luxury good but that very well appears to be what's happening here.

Android Q is only going to be available to very new devices. Whereas, the vast majority of Android users have older devices and won't be getting over-the-air updates that actually patch this vulnerability. So, unless you're willing to drop $800 on a new device, you're probably going to have apps still exploiting these vulnerabilities on your phone. So, that's it.

[APPLAUSE]

LERONE BANKS: Thank you, Serge. And, finally, we'll have a presentation from the Yixin Zou, who will talk about breach notices and some ideas about enhancing them.

YIXIN ZOU: Hi, everyone. Thanks for the intro. Before I start, I'd like to acknowledge the contribution from my colleagues at the University of Michigan, my advisor Florian Schuab, who is sitting right there, and the funding from the Mozilla Corporation.

So, we all know that data breaches are security incidents that compromises the sensitive and confidential information of individuals. Concerningly, data breaches are on the rise. Look at this figure-- the number of data breaches per year has increased from 157 in 2005 to over 1.2K in 2018, resulting in almost 450 million exposed records in just one single year. There are several potential harms of data breaches, such as data leaked to the dark web and being used to conduct phishing attacks. All of this can lead to identity theft, which result in substantial financial loss and emotional distress to victims.

Many laws in the US now require data breach companies who suffer data breach to send notifications directly to affected consumers. For example, all 50 states now have enacted their own data breach notification laws, in addition to a few industry specific laws, such as HIPAA

for health institutions. However, there is no consensus of when to notify consumers, what content should be included, et cetera, resulting in large inconsistencies in between.

In contrast to regulatory requirements, we see empirical data showing that consumers are not taking enough actions. In a 2014 national survey, 32% of respondents reported their first reaction to a breach notification is to ignore it and do nothing. Not surprisingly, consumers have bad security practices that leave them vulnerable to data breach risks as well.

In a 2017 survey, 56% of respondents reported they used the same password for multiple accounts, which means if one is exposed password is exploited it will lead to chain reactions to the other accounts as well. So, now we all know that data breaches pose significant security risks. Data breach notifications while mandated by laws do not trigger customer reactions effectively.

This brings to question-- how to make data breach notifications useful, which I will explain in the two studies I'll present today. Our first study studied consumer reactions in the context of the 2017 Equifax data breach. You probably remember Equifax is one of the big three US credit bureaus. And this breach of fact it almost half of the US population-- affected their names, addresses, social security numbers, along with other sensitive information.

Our first research question is how did consumers perceive the risks of Equifax data breach. The second-- what protective actions did customers take in response and what are the reasons behind action or inaction? For our methods, we conducted 24 semi-structured in-person interviews between January and February 2018. We recruited participants using social media and email list and also use a screening survey to diversify demographics.

And then after transcribing the interviews, we use thematic coding for analysis. Along with prior work, we found most participants took little actions despite high concern. 20 out of the 24 participants were aware of the breach. Identity theft and privacy invasion were conceived as two primary risks. However, 14 participants did not take any protective measures.

We further worked participants through a list of suggested actions by the FTC. And then the majority of participants reported they were either unaware of the actions or avoided taking them intentionally. Even for those who took actions, most choose reactive approaches that do not fundamentally rule out the possible risks. Only four participants took proactive measures and more stronger measures, such as freezing their credit reports.

We further provided novel insights of why this inaction might be the case. Many of which are participants on cognitive and behavioral biases. For example, several participants exhibited optimism bias, saying, why would they go after me if there are rich people out there? Other participants showed a general tendency to delay actions until harms have occurred.

Moreover, even for participants who took actions, they may have a false sense of security that discouraged them from taking other actions, such as keeping an eye on their credit report after freezing their credit-- keeping an eye on their credit reports after credit freezes. There are also extrinsic factors that motivate actions or inaction.

Source of advice, for example, is a prominent one. Whereby, advice from family members and colleagues were more effective at triggering actions, compared to news media, which

informed participants of the event but did not do much beyond that. Cost is another big factor.

For example, credit freezes were not free back then. And some participants mentioned this as a reason why they wouldn't use it. Finally, many participants misinterpreted the functionality of certain measures, such as viewing fraud alerts as alerts sent from banks when fraudulent activities happen in contrast to its real purpose as a red flag on one's credit report to signal identity theft risks.

Our first study indicates that there might be issues in current data breach notifications that impede customers to develop a correct understanding of the protective measures available. Our second study is a systematic empirical analysis to further unpack those potential issues. We looked at issues regarding a notifications readability, risk communication techniques, structure, and format, as well as how recommended actions are presented.

We sample 161 notifications during the first half of 2018 from Maryland Attorney General's website, which requires companies to upload their data breach notifications according to the state law. We collected quantitative metrics for the readability analysis and also qualitatively coded notifications for diverse communication and presentation practices. We find that data breach notification are indeed hard to read.

Using a Flesch reading e-score as a measurement metric, most notifications receive a score between 30 and 60, indicating they're difficult or fairly difficult to read. Using the word counts of notifications and an estimated 250 words per minute speed for average adults, we calculated that the estimated reading time for a notification is six minutes. This may be, OK, compared to the notoriously long privacy policies, but consider things more common in our daily life, such as a news article, which takes two minutes and an email which takes no more than 20 seconds. The six minute paired with the need for advanced reading skills still creates a considerable burden for consumers.

It's also possible that companies use techniques to downplay potential risks. For example, 70% used hash terms when describing the likelihood of the recipient being affected, saying, "We recently identified and addressed a security incident that may have involved your personal information." Or they can use a low evidence claim when describing the possibility of exposed data being misused in 40% of our sample. They may say, "We're not aware of any fraud or misuse of your information as a result of this incident." While those statements might be true, they're still misleading by making customers think there are no future risk and no actions are needed. A better practice will be at least adding a sentence like, "Still, we urge you to take this action out of precaution."

Moreover, there might be a choice overload problem for recommending too many actions. Eight is the median number of suggested actions in our sample. And, notably, important measures, such as credit freeze was first mentioned in the appendix instead of main text by 73% of notifications that mentioned it. To make things worse, actions were often buried in landslide paragraphs instead of being highlighted effectively.

In this example, one has to read very carefully to know the first paragraph talks about fraud alert and the second one talks about credit freeze. Still in the same example, there's very little guidance or indication of which one is more effective between these two measures and thus

should be prioritized. This may be common knowledge for experts that credit freeze is more preventative but general consumers may not know that.

We also see evidence showing that companies are leveraging existing templates to create notifications. For example, 94% of notifications that use headings follow the exact wording suggestions in California law. This is promising, because headings help parsing things out and guiding the reader's attention.

On the other hand, there are places where companies need to make adjustment to existing templates but fail to do so. For example, many notifications appended a long list of contact information for different state attorney generals. For a Maryland resident, this might not be needed and decreases readability significantly. The summary of the two studies that I presented today.

First, customers do not react to data breach notifications for various reasons, both due to their own heuristics and behavioral biases and due to the issues with notifications themselves. And second, in order to make data breach notifications more effective and motivating customers they should take actions, we need to fix these uncovered issues. I want to highlight three recommendations that are particularly targeted at public policy. The first one is to incorporate readability testing based on standardized metrics into regulations.

Compared to the current regulation that vaguely says the notifications shall be written in plain language, there should be a better way, such as using readability metrics, to make him more specific and actionable for companies to pursue. In fact, such readability metric-based practice has already been adopted by the insurance industry for regulating health insurance policies. For example, this is a snapshot of Rhode Island's state regulation.

The second recommendation is to provide concrete guidelines of not only when customers need to be notified and what content needs to be included, but also how in information should be presented. Using this bad example, again, we can easily brainstorm some potential improvements. For example, place credit freeze on top of fraud alert to indicates its importance. And second explain why is it important for the recipient to do so by, for example, showing personal salaries or making connections to the types of breached information rather than dumping all definitions and instructions altogether . And third, visuals should be used to highlight the key information, such as which paragraph is about what protective measures.

Our last recommendation is to leverage the influence of templates to advocate positive changes. FTC, for example, has provided a model privacy form for financial institutions to comply with the GLBA. Similarly, we can start by designing templates that encompasses suggested best practices and add a template to regulations. And we've already seen evidence showing that companies are using existing templates.

Therefore, it's promising that if we can have a federal level breach notification law with a good template validated by user testing, then companies are likely to adopt this best practices since they don't want to reinvent the wheel either. So, this is all for my presentation today. I'd like to thank again for the funding and support the University of Michigan and the Mozilla Corporation, as well as you listening.

[APPLAUSE]

ANDREA ARIAS: Thank you, Yixin. [INAUDIBLE]. OK, we'll now turn to questions. We'll start off with some observations and questions from us. But I encourage you all to go ahead and submit questions either through the comment cards-- raise your hand and one of our colleagues will come and give you one-- or if you're watching through the webcast or if you just simply don't want to write it down a comment card, you want to just tweet it to us. Go ahead and do so using the @FTC #PrivacyCon19.

So, it seems to me that you are all striving to answer some of the same basic questions, which is what issues plague the security of our ecosystem and what can we do about it, whether it's risk assessments and risk analysis and vulnerability analysis, whether it's actual leaks-- we're particularly talking about them in the app system-- or even just how do we notify consumers through the breach notification process. So, we're kind of thinking about a few questions which is I think the goal that we're going to be try to answer today which is, how can we effectively identify and rectify data insecurity? What are the incentives to invest in data security and are they enough?

What's the best way to inform stakeholders-- whether it's security personnel, executive board, cyber insurance, card issuers, regulators, or even consumers as we've learned through the breach notification process-- of the state of security at a particular company? And, finally, what regulatory and enforcement approaches are working and how can they be improved? With that thinking in mind, let's turn to our questions.

And I'd love to start with both Serge and Elleen because you both found that Android app developers are accessing user's private data maybe without consent or using side and covert channels, undermining users' privacy. So, if permissions exist, why do apps feel the need to bypass the permissions to obtain this data ? And if so, is there an inherent problem in our mobile permission system? Either of you want to say?

SERGE EGALMAN: Sure. I mean, to answer the question, why would app developers want to go around that? I mean, that's like asking why would someone want to steal something when they could just buy it, right? I mean, I think the fundamental issue though that has kind of-- the common thread among all the things that have been talked about today is that, fundamentally, consumers have very few tools and cues that they can use to reasonably control their privacy and make decisions about it. So, regard to permissions and Android, if app developers can just circumvent the system, then asking consumers for permission is relatively meaningless.

Because even if you decline, it's possible that the app is still accessing that data anyway. Fundamentally, I think that this is a policy issue and that enforcement needs to happen to go after some of these deceptive practices. I think one thing though is that with most of this-- so, certainly, in my work this appears to mostly be occurring in third-party SDKs that get bundled into apps. And I suspect that most of the time, it's probably likely that the individual app developers probably don't have a good grasp of what's going on when they bundle third-party code in their apps.

And so, there needs to be better guidance for app developers on behalf of the platforms as well as the SDK developers in terms of documentation for the privacy behaviors of the SDKs that they might be bundling. But then at the end of the day, when there's something that's relatively egregious, I think there should be enforcement action. So, the example I showed of the app first checking to see whether it has the permission.

And then if it doesn't, jumping to you know the exploit code-- that's not an accident. I mean, the developer made the choice to do that. But, again, I don't think that consumers should be really burdened with having to figure this stuff out on their own. I used to say sarcastically when people would ask me what can consumers do to protect themselves better, and I would say just do what we do, which is implement platform instrumentation, bespoke network monitoring code, and read through the assembly of the apps that you're running. Obviously, that's absurd. But at the end of the day, that's kind of where we're at with what is expected of consumers if they're to actually understand what's happening and make decisions about it.

SASHA ROMANOSKY: So, there are some efforts to develop a software build as material by the Department of Commerce, NTIA, and Alan Friedman especially that speak to exactly that point. Currently, we have no understanding as users-- apparently even as developers-- what libraries were including in the software that we build, whether it's web applications or mobile apps or whatever. And so, one possible solution to that is to develop this requirement-- maybe guidance, maybe a requirement-- to help disclose these libraries that are included in order to better understand. Now, to the extent that that transparency-- that kind of transparency-- over any other kind of transparency will magically help is to be determined but there could be a real solution there.

ANDREA ARIAS: Elleen?

ELLEEN PAN: Yes, I guess just to like add on. For the stuff that we found, there wasn't even a permission that could be asked for. So, in those cases, it's like there's nothing to bypass. Because you can just have access to the screen if you wanted.

LERONE BANKS: Elleen, I had a question related to some of what you saw. Did you see for the apps that did screen recording that they would record when the app itself was in the background, and then, they would be doing screen recording of a different app that was in the foreground?

ELLEEN PAN: So that type of behavior is actually protected by a permission. And the one that we noticed was not that. So, it was only in the app itself. But it's still going to a third party. And it's still recording things that could have PII in it.

LERONE BANKS: And so, based on that, I guess the other panelists could give your opinion too since you are users of these devices as well. But does that suggest that there is a need for a screen recording or screen capture permission or a different one or an enhancement to the existing one? I guess, Elleen first since it was kind of a part of your work, but anybody else that has--

ELLEEN PAN: Yes, I guess, like just thinking about how the main app and the third party permission should be separated if there was a permission for this. It's totally valid for an app developer to have access to their own app provided they like hide sensitive information and stuff. But in terms of third parties, users are completely unaware of this type of behavior happening.

ANDREA ARIAS: We have a question from the audience. and I think it relates again to Serge's and Elleen's work. You both focused on Android apps, but they want to know does Apple have the same side route vulnerable or issues that you guys saw in the Android area.

ELLEEN PAN: For us, the library that we saw, specifically Appsee, they also have an iOS SDK. So, it's likely happening there as well as it's their only business model.

SERGE EGALMAN: Yes, the main reason why we pick on Android is just because it's open source which means it's relatively-- it's more straightforward for us to add instrumentation into it and then be able to test off the shelf apps. iOS is closed source, so we just don't have the capability of adding the same level of instrumentation. And also all of the apps are encrypted. And so unless you have a jail broken device, which is instrumented to decrypt those apps, it's not a straightforward process to get the same insights on iOS.

That said, we've done some work just looking at the network flows coming from iOS. And by and large, the same third parties are present in both apps. Whether the same security vulnerabilities exist, who knows.

LERONE BANKS: Yes, so I guess on that iOS point-- and this is just a general perspective for really all of you that have looked at apps and vulnerabilities. To what degree do you think that Android and Google as a part of that sort of benefit from the more open nature of their operating system? Is it something that maybe regulators should either push for additional hooks in other platforms or an idea to push for hooks in platforms that would make it easier to do this type of analysis? Or is the environment taking care of itself?

SERGE EGALMAN: I don't have a good answer to that actually. You can take some time and think about it if you want.

SASHA ROMANOSKY: So, there is a bit of work. It is an old question of what is more secure, open source or closed source software? And there is a fellow at Boston University, Sam, who has done some work on that and found that there was a little bit of improvement in open source software. Although, there is some nuance, which seems to be pervasive in our field that it's not universally true and it's not substantially true that one has an improvement over the other. But it's a great question. And the more effort-- the more analysis that can be done in that area, I think all the better. And that could help you all with your efforts to try and promote or facilitate, again, more transparency with different kinds of software development.

SERGE EGALMAN: To their credit, one of the big differences that we've seen between the platforms is just that Apple does a lot more vetting of the apps that are distributed in their app store. Whereas, Google it's largely a free-for-all. So, they analyze apps in Google Play for malware. But beyond that in terms of compliance with policies, there's not really any of that. Whereas, there seems to be on iOS.

That said, the vetting process on iOS is a black box for those outside of Apple. So, it's not clear what exactly they look for. But certainly with regard to policies around transmitting various persistent identifiers, that seems to actually be enforced under iOS insofar as developer wanting to put their app in the app store. Whereas, it's not in Google Play.

LERONE BANKS: [INAUDIBLE]. I guess, just one other follow up-- it sounds like there are a few different options for future approaches for regulators and platforms to consider. So, one would be to add in additional hooks that make analysis and observation more easy within certain platforms. But another one, just based on some of your responses, relates to app store scrutiny and how much effort or engagement the app stores themselves or the platforms put on monitoring the practices of particular apps.

And this is for any of the panelists-- do you have a perspective about which approach might be better? And "better" is pretty vague in the sense of more feasible to actually be implemented by the platforms, faster, more efficient to actually do or even realistic. So, you can decide in your own terms sort of what better means. But of those choices, based on your experience, does there seem to be one that might be better in some way? These are just thoughts.

ANDREA ARIAS: You stumped them.

LERONE BANKS: That's not my goal. it's really trying to.

SASHA ROMANOSKY: To the extent that you can drive accountability by those platforms, that would certainly help. Now, how possible that is? I don't know. I guess that's your challenge. But if there were accountability by Android, by iOS at vetting these apps, then, yes, presumably you would think that would have an effect. And, actually, if that were to happen but that would be a great situation for empirical research. So, if there was any way you can affect that, then I'm sure there'll be lots of people who would love to take that on to see if there actually is an effect of more secure apps or at least less privacy invasive apps being developed and being uploaded after the fact.

SERGE EGALMAN: I think forcing all app developers to share their source code with the government is probably going to be a uphill battle. That said, it's not it's not as crazy as that seems if you-- I mean, there are analogs to that. So, for instance, with telephony, there's CALEA, where telecom providers are required to have basically back doors so that they can service wiretaps. And so maybe you can think of some sort of similar thing, where the platform has some way of auditing the collection of personal data. But that seems like that's very far off.

I mean, honestly, I think that a big improvement would just be having the platforms proactively enforce their own policies. So, one thing that we've been seeing a lot of is both Apple and Google have policies about all the advertising and tracking that's done needs to be done using the resettable advertising ID as opposed to hardware based persistent identifiers that can't be reset. And then there's a privacy settings interface, where users can go to and periodically reset that identifier, which is akin to clearing your cookies in the web browser.

The problem is if those identifiers are sent alongside other persistent identifiers that can't be reset, that just totally negates the privacy preserving value of doing it that way. Both platforms have the same policies that everyone who is doing advertising and tracking should only use that advertising ID. Whereas, we've seen with Android apps, the majority of them appear to be sending the ad ID alongside other persistent identifiers violating Google's policies. And that's just because the policy is just totally unenforced. I think that enforcing the policies that they're representing to consumers-- they represent to consumers if they have these policies to protect consumer privacy, but then kind of turned a blind eye to any violations or even checking to see whether they're violations, I think that's a pretty big issue. And if just that alone were solved, we'd be in a much better place.

ANDREA ARIAS: Yixin, I'm going to turn to you. There's a question from the audience, which just says, do you prefer multiple strict state level regulations on breach notifications or a federal maybe-not-so-strict regulation that would preempt state regulations so as to provide clear guidance to the ecosystem? And then I'm going through a bonus question at you, which

is since you've reviewed so many breach notifications, have you found maybe an example of a data breach notification that meets all your policy proposals?

YIXIN ZOU: Cool, so for the first question, I will say it's tough to answer because it's restraining me to two very bad choices-- like most users. Yes, I will definitely advocate for strict federal level data breach notification law. I think we should be aiming for the higher standard. And if that cannot be achieved, then we make compromises.

But the ideal case is if a law like GDPR can be implemented at the European Union level, then there are certain challenges. But I think those challenges can be conquered. And there are multiple calls for a federal level breach notification law in the US. So, I think we definitely have motivations here.

Between the two bad choices, personally, I would prefer a federal level with less restriction. Because still if there's] multiple state strict policies, there are still inconsistencies in between. And when you have a breach that effects so many state residents, each state's residents have different levels of breach notifications with various levels of compliance to the strict laws. So, that's a problem. But, eventually, I think a strict federal level law will probably make the most sense.

And to your second question, I will say very, very few would maybe meet all of my expectations. There are a few that I can remember, probably, were they use very short but effective text. And then they use visuals to show the recommended actions in the very appealing and attention-catchy way. But I will say that number is probably like less than 5%, so very concerning. And on my own research effort, I'm thinking about ways that can come up with potential templates for incorporating all those expectations and then how can we put this kind of template to a wider range of audience.

LERONE BANKS: Yixing, one follow up to that and this is from Twitter, given a low response rate that you observed, do you think that organizations should be more proactive in providing credit monitoring since they know that consumers won't necessarily respond?

YIXIN ZOU: Yes, that's a good question. I think definitely it's reasonable to assume there's consumer fatigue for data breach notifications, given that data breaches are on the rise and customers receive tons of notifications on a regular basis. So, they may ignore that. But I think is still important for organizations to provide such services, given that customers are-- it's their right to have these kind of protections and compensations for their loss in a data breach incident. And an effort should be made to, for example, how to present those measures that they deserve in a more readable and actionable way to engage them into this opt-in process or even opt them in the protections automatically.

LERONE BANKS: I'm sorry. I kind of misread the question. So, your answer totally makes sense but the part I left off that is relevant in the question is the question emphasizes whether or not companies should be required to offer credit monitoring without requiring necessarily action on the user's part or on the consumer's part. So, now, though the structure is that a consumer receives a breach notice. And then, typically, they have to take some action. They have to go to a separate site or give their information to some other service. And so the question, I think, is trying to get at whether or not that stuff should-- to whatever degree is actually possible, whether those services should be offered automatically to consumers that are implicated in a breach.

YIXIN ZOU: So, are you asking is it reasonable for us to mandate companies to--

LERONE BANKS: --do it automatically?

YIXIN ZOU: Yes, I think so. But I can expect companies will say there are challenges. I do not see what challenges are out there. Maybe other people can offer perspectives. But I think it will be similar to GDPR privacy by default. This is security protection by default.

LERONE BANKS: And I'd welcome some other comments from the other panelists. I mean, we're all consumers in this case. And we have perspectives on sort of whether automatic credit monitoring services would be bad for us as individuals. So, feel free to chime in.

SERGE EGALMAN: Not about this but something that Sasha said I actually wrote down a comment. You mentioned that you think firms are good at finding vulnerabilities but not good at fixing, which actually parallels a recent finding. I have a student who is interested in looking at the vulnerability discovery process. And she's been doing a series of over 50 interviews now with various stakeholders, all the way from independent bug bounty hunters all the way up to management at large organizations who are responsible for their organization's approach.

And one of the most astounding things, I thought, that she found consistently when talking with management was that a lot of companies have made conscious decisions to not have any sort of vulnerability discovery process in their organization just because they don't have the resources to deal with fixing anything that might be found, which is--

SASHA ROMANOSKY: --bonkers. Yes, I mean that makes sense. I mean, look if you're running a medium-large firm, you're going to do your vulnerability scan. You're going to find hundreds of thousands of vulnerabilities out there. How do you possibly make sense of that? Maybe you're just struggling to keep the business running as it is.

And you're human, you're susceptible to distraction. And you have limitations and constraints. So, I could totally see that. There is one advantage of them that they can act as forcing functions. So, it makes that whole vulnerability issue a public thing. And so for firms that do have an interest in maintaining reputation, it could actually help them stand up a proper team to fix these vulnerabilities. So, I think there is a value there.

To your point though of requiring or facilitating-- I don't know-- mandating companies to provide credit monitoring automatically. It sounds like a good idea. I could see it being hugely inefficient though. Because what do you do with people that already have the credit monitoring? Now, they have two credit monitoring. And surely there's a cost to that. If greater than 0-- maybe it's epsilon, it's greater than 0. And so it's not clear what the marginal benefit of that is. So, I don't necess-- it sounds like a good idea. I think it would just be too inefficient.

LERONE BANKS: One other, I guess, question about the breach notice study. It's not mentioned in your paper, Yixing, but I'm just curious if during the interviews whether or not any of the participants offered alternatives to some of the options that you listed as responses to breach notices-- so, like checking the website or setting up for free credit monitoring. And, specifically, what I'm thinking about is whether-- all of those actions seem to be focused on trying to prevent misuse of the loss of data, right? My question is really whether or not any

other interviewees or anyone has thought that much about changing the relationship with the breached entity.

So, nobody says after they receive a breach notice, I'm going to go back to a company and reduce the amount of information that I have with them or, in some instances, even terminate the relationship with the breached entity. I know that's not feasible in every scenario. But my question is just to what degree any other interviewees may have offered alternatives, whether or not you think any of those other alternatives might be interesting--

YIXIN ZOU: Well, that's a good point. So, unfortunately, no participants in that study explicitly mentioned they would like to terminate relationship with Equifax. And I think Equifax is a special case, given that this is a credit bureau that just proactively collect every US-- people who live in the US, regardless if you're citizen or not, they have information about you. They collect your information through like third parties, like banks and other creditors that you interact with without necessarily obtaining your consent.

So, in that way, they don't have the choice to terminate the relationship. With that said, I do remember there's like a survey by RAND Corporation a few years ago when they ask, do you terminate relationship with a company that suffered a data breach? And I vaguely remember the number is 11%, so very low. And my assumption is that just most people when dealing with this kind of data breach incidents, they start with a very reactive approach. Their focus will be how to stop the misused data rather than doing it in a proactive way, thinking about what behavior changes I can do to prevent those companies from collecting my data in the first place.

LERONE BANKS: Do you think that it would be worthwhile-- and any of the other panelists feel free to chime in too-- to offer other options to consumers that are victims of a data breach that allow them to change in some way the relationship that they had with the breached entity? So, whether that means-- and I'm totally making this part up. But whether it contained within the breach notice, there is some link that allows you to go back to the entity and maybe either see the data that you have or delete some of that data. And I know Equifax is a special case, but you can think of other types of breaches like maybe Target's from a while back, where a consumer can maybe change some of the information or delete some of the information. Do you think that that process would be something worthwhile to integrate in the breach notification process at some point?

YIXIN ZOU: I can start. So, personally, I think that's some very good advice to think about. I do have the impression for current breach notifications, they focus on two aspects. One is the financial related, especially if it compromises social security numbers or other sensitive information like this to enrolling credit freeze, fraud alert, anything related to preventing yourself from identity theft. And the other thrust is the account protection service and online service, then they ask you to change passwords or other related behaviors to secure your credentials.

But I see less recommendations about, for example, like more to the privacy-focused recommendations about reducing the amount of disclosed data or even a simple action like delete your data or review your privacy settings with this company. It's still a burden placed on customers. But if they do that-- and I think more research is needed to see if this kind of proactive actions implemented by customers can reduce the harm caused by data breaches for the long term.

ANDREA ARIAS: Sasha, I have a question for you. So, in your model, you talk a lot-- and maybe because I've had the pleasure of reading the paper. And I know a little more than the audience does. But, basically, in the model, it talks about how it focuses on threats-- the likelihood that a vulnerability will be exploited. But you do say that you don't consider other factors, such as the strength of security controls employed by the target organizations or the value of the asset of the data or the actual impact a successful exploit would have. So, if you could maybe take your model and include those-- have you guys considered taking your model and including those factors? And if so, what do you anticipate maybe the results would be?

SASHA ROMANOSKY: So, if we did all of that, we would have effectively what is CVSS, this vulnerability scoring system, in its full blown capacity. So, that standard was started 2003-ish and was built off of three components. It's trying to assess the characteristic-- or at least describe the characteristics of the vulnerability itself, describe characteristics of how an exploit may be developed and used, and describe characteristics of the firm and the firm's controls.

And so, its severity. It's a little bit of threat. And it's an environmental component. And together, all of that should represent the risk. And so, we built that standard kind of with the hope that it would actually become a risk scoring standard for vulnerabilities. At the end of the day, what really took off was the severity part of that. And so, it's done a very good job over these years of describing the severity that a vulnerability would cause to your organization if that were to be exploited.

The other stuff was left-- I mean, it still exists, but it's much more difficult. So, the idea being that any firm or every firm would take that vulnerability, they would understand the severity part. And then they would integrate and try and understand and incorporate all of their controls to then reduce or increase the severity in order to represent the risk to the firm. But like I said, that didn't really take off because it takes a lot of work.

Each vulnerability that comes in that's disclosed needs to go through that process. So, that's a very labor intensive kind of thing. And it's also very difficult. Because there's a lot of measurement error. There's lots of subjectivity to try and to understand, OK, still, what do you need to do with that? And so, that's kind of where we are now.

It exists. Some people still use it, but it hasn't really worked-- it hasn't really taken off. With this threat scoring system, right now, what we're trying to do is really take very objective measures about the vulnerability and about the ecosystem to try and understand what is the probability that this vulnerability will be exploited in the wild. And so, you're right that exactly we don't take into account other controls by the firm, which then would really give you what you actually want which is an understanding of risk.

But then we're back to the same spot that we were before of, now, each firm needs to go through that step over and over again. And that's a lot of effort. And it's an unsolved problem on our end. There are different software applications that try and do all of that.

They ingest your firewall rules, your vulnerability scan data, your router ACLs, and try and present you this kind of ranked order of vulnerabilities that also incorporate user driven assessments of the value of your assets and kind of gives you that risk. And so there are packages that do that and they do that fine. I don't think we're going to be able to solve that

just yet, but who knows? Maybe what we develop in our scoring system, firms can then incorporate into their practices in an automated fashion to answer the questions that we really want at the end of the day, which is about risk.

ANDREA ARIAS: Great. So, we are almost out of time. But I do want to do a very, very quick wrap up. And then Lerone's going to give us some brief remarks. So, I hope that you all, like me, think that one of the biggest benefits of PrivacyCon is that it brings together the best and the brightest, who are all working to understand the same issues. And I think we've done that here today.

So, I hope this session and honestly all sessions today facilitate you learning from and building upon each other's work. So, we also hope we can continue to benefit from your insights about how best to protect consumers' privacy and data security. So, we're grateful for you all to coming here to share your work and your thoughts with us today. So, let's give our presenters round of applause, please.

[APPLAUSE]

And with that, I'll turn it over to Lerone Banks, who's going to give us some brief closing thoughts.

LERONE BANKS: First, I want to thank everyone who has paid attention to this event today, whether you're in the audience with us today or watching us online somewhere. One final thought that I want to give is that there has been some talk about the value of technologists or people with technical skills being involved and working with the FTC. And I really hope sincerely today serves as a emphatic answer to that question that the people that presented today represent a small fraction of all the great work that technologists or people with technical backgrounds are working on today.

And I hope that we can continue to build on the knowledge and energy that's been generated today to figure out how the FTC and federal agencies in general can work better with technologists. So, with that, all of the papers from today's presentations will be available on the FTC website. I hope you enjoyed today's event as much as I did. And we'll see you next year.

ANDREA ARIAS: Thank you everyone.

[APPLAUSE]

LERONE BANKS: That concludes everything for today. Oh, I'm sorry. I forgot one very, very, very important thing. A lot of people worked hard to put this event together today. But in particular, Andy and Jamie-- they spearheaded everything. They herded all of the cats, which I don't know what a big amount of cats are called. But whatever that word is, Andy and Jamie were very instrumental in bringing this together. And so, I hope we can all take a little time to thank them as well. And thank you very much again.

[APPLAUSE]

[MUSIC PLAYING]