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

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 LibRadaelostud badifferek4s(6.8 (the funct5.9 5 (7dr(6.8 (l5f5uc7% of t7.9 (ed)0.9 atised offik31.6

Sctrariesaurkudyo, inJwh1 Tw 0 i6.2i5Tw]TJT*(1.1n (c7.7(1.2 9eren.so t7.9 m-5.31 5001 Th (c7.7(1.58 Tw 0 -1)

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

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 compsagrti.2 ( )2(.)1 h.a.5 7n.1 (-)4.2 (ti.2 (5 (v(re cop0.9 (m)11p5.3i( )2(11i(6.1

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

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

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 th

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.

GARRETT JOHNSON: Well, thank you very much

think is more attributable to the GDPR. So, using this strategy, we see a across the board

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.

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

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

we learn 2 h.2701496e r.27.1.1i6.1.1v-0..9 (cy-0..9 (regula3 TTj Tw1.1ion-0..9 (0008 Tcthat48 Td[()-4

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. 5ei (n g)5.4S'll just brief

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 tei (n g0 -1.148 TD[(start u

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

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 m -1.148 TE3 (an(what we7.6 ( )10.6 (m6o(u 5.3any)11 (17 (what we 7 (ion.Tc 0.0004 Tw.Tc.0005 e (i

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 imbalancisr us. s w11.6 n4.5 (t)6.1.3 (our dat-)]4.7 (5 (n

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.

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

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

Our work was covered in the press, and it was

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.

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 com.m.t da7gp

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 requi

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

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

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.