FTC PrivacyCon 2017
January 12, 2017
Segment 2
Transcript

JUSTIN BROOKMAN: With the second session now. So I am Justin Brookman. I am the Policy Director of our Office of Technology Research and Investigation, OTEH, here at the FTC. This session is going to focus on mobile privacy issues. This obviously has been an area where the FTC has been active for quite some time, whether it's policy guidance, like our report on mobile disclosures. Whether it's tools. We have a tool for mobile health app developers, to test to see what sort of health privacy laws might be implicated. And enforcement work, as well, starting with the Golden Shores case. But also, more recently, cases like InMoby, Turn.

And we certainly recognize these mobile devices are, obviously, incredibly useful and really amazing. I mean, right? We're all kind of obsessed with our phones, maybe a little too much, but it's because they're incredibly practical, and useful, and functional devices. At the same time, we recognize there are these heightened privacy concerns, that we've been aware of for some time. But I think we're still grappling with in a lot of ways. Obviously mobile applications and libraries have access to information that's maybe a little bit harder to get in other contexts, like the web. Access to sensitive data, like geolocation, sensors like microphones, and cameras.

Obviously we've seen the platforms, kind of, evolving to try to address some of these concerns, over time. Into the development of specific advertising identifiers. Even more recently, you've seen Apple starts to move away from device identifiers for people who try to opt out of targeting. You've seen Google try to adjust, and moved more to a first time a permission model, for some

you'd like to ask the panelists. You can also ask on Twitter, we have question cards that will be sent up here. I may try to follow it as well. And with that, let's get going. Dave Choffnes, please.

DAVE CHOFFNES: All right thanks for the introduction. Let's see if this works. Great. So today I'm going to be talking about Recon, a tool we built for revealing, and controlling personal information leaks from mobile network traffic. And just to get started I want to make this a little bit interactive, wake you guys up. So how many in the room have used your mobile device to access the internet today? Just raise your hands. So for those on the web, pretty much everyone. OK. Now, how many of you have used it just since this session started? In just the past couple of minutes. We have a lot of honest people here. About half the room has their hands up.

So I mean, the point I'm trying to make here is, obviously mobile devices have become essential, and we're addicted to them

So high level, one question you might ask, I said machine learning. There's just some technology I'm throwing at this. Does it work? In our lab experiments we found it's very accurate, with very few false positives, and false negatives. So in case you're wondering, generally you don't see too much bad information, and the system learns over time. We've also given this out to users. And we've had almost 400 users who have enrolled in our user study. As part of this we surveyed users at an early stage. They found it useful. In fact, some modified their behavior based on the information they learned.

On top of that, we found over 27,000 cases of information being leaked by various apps. And there's been numerous cases that are suspicious, or in some cases actually just simply dangerous. So I'll focus on some of those, which is, in the process of trying to understand privacy, we saw that passwords are being exposed in plain text. Or sometimes, they were encrypted, but exposed to third parties that never should've gotten them. So far we've identified-- I should now update this to 26 apps that have exposed passwords, because we found another one last Friday, in an app that is used by-- well, they claim over 100 million installs.

We do responsible disclosure, which is why I'm not going to name them right now. So we wait until they fix the problem, before we go public. And we've got a variety of responses that are somewhat interesting from developers. And many, particularly at the major companies, will act quickly, but some don't necessarily understand that this is a huge security problem. Some actually did it by design. Some aren't even able to fix the problem, because they don't have access to the source code, and the vendor that they used doesn't exist anymore. So these are some of the problems that we face. These are persistent problems. So it's not a matter of just fixing privacy and security concerns at a certain moment in time. You have to keep monitoring. And you have to keep being vigilant. And you have to have systems that react, even when the app vendors won't be able to do it themselves.

So if any of you are interested in some of the results, we only publish the ones where we're 100% certain which app is actually responsible for leaking information. You can find information about this on our website. We also have a version that looks at websites that are leaking information. It is not our focus, but it's something that we also see as part of this study. And so to wrap up, with this project, what we're trying to do is improve transparency, and control, over personal information. So what we do is we learn what information is being leaked. We use crowdsourcing to determine if were correct. And also give us hints as to what matters to consumers. And we allow those consumers to block or change what's leaked. This is an ongoing project. You're about to hear from Narseo about Lumen. And this is something that we are talking about integrating into that environment. You could also build this into home routers. And we'd also like to apply this analysis to IOT devices, which we just learned about in the previous session, to understand what personal information is being exposed by those, as well.

So before I wrap up, I just want to thank my collaborators. In particular my Ph.D. Student Jing Jing Ren, who is behind most of the work. And if any of you are interested in learning more about Recon, you can visit this website right here, where you can also sign up to participate in our study, and use the system yourself. That's it.

JUSTIN BROOKMAN: Narseo. NARSEO VALLINA-RODRIGUEZ: I hope that you can hear it. So I am Narseo Vallina--Rodriguez, and I'm going to talk about our ongoing efforts to illuminate the [INAUDIBLE] system, within the Lumen Privacy Monitor. This works is done in collaboration with a lot of colleagues from the International Computer Science Institute. That goes from Verne Jackson, to Mark Calmet, Kristin Cliby, Sir Cheoman Edwin Driaz, and also Primal. And other colleagues at UMass, and Stony Brook University.

So whenever we run a mobile application, we know that they are accessing certain pieces of information, which they can use later to create an accurate profile about our persona. And we know that they are accessing this information because we are granting them permissions to access these pieces of data. The problem is that most users will believe that this information is only shared with the application developer. But the truth is that this information is also shared with a large number of third party services for analytics, and advertising purposes. Unfortunately, as opposed to the desktop context., we cannot rely on existing ad-blockers, because those are specifically targeting web apps. And in the case of the mobile applications, advertisement downloads are integrated in the app.

So in this project, we have three specific goals. The first one is to define the third party ecosystem that exists on mobile systems. Then we want to evaluate their impact that they have to use their privacy. And finally we want to promote more transparency, by releasing the data, and bhGnfe tir6(o)P-12(l)-4(17.05u)]TJ 25.76(y)20-2(,[(b))-4(s)-5(ea15.9(y)-5(ea-4(ec)-1.15 en[(b) a f)3(1r)-11

The challenge that, first we have to tackle is, identify the domains that are related with first party tracking services, or with third party tracking get services. And for that we represent the interactions between mobile applications, and domains as separate. And here you can see two examples, accuweather.com, which is a well-known weather service, and [INAUDIBLE], which is an analytics service. And the basic heuristic will be analyzing, or considering a domain as a third party service if at least than one application is talking to it. But, as you can see in those examples, the Accuweather app, is also accessing accuweather.com, in addition to the HTC weather wizard. So how can we disti

SEBASTIAN ZIMMECK: Thank you, Justin. And it's a pleasure to be here. I'm talking about the automated analysis of privacy requirements for mobile apps. And I was fortunate to work with a shipload of great collaborators on this. And all of the people who worked on this were, at some point, part of the usable privacy policy project, which was funded by NSF, DARPA, and the Air Force Research Laboratory. So when you're using a phone many different types of data are sent to first, or third parties. For example, device IDs, location information, e-mail addresses. And these types of information should be described in a privacy policy, you know, what is collected, what is shared, how long are these data retained. And the idea of this project is to look at the privacy policy on one side, and analyze what is stated there, and on the other side, look at the apps, and see whether what is said in the privacy policy actually is happening in the apps. And we call that a privacy requirement compliance.

So we need to look at both sides. The privacy policies, as well as the mobile apps. And because we want to make this automatic, as automatic as possible, we decided to analyze the privacy policies using machine learning. Essentially we are looking at the text fragments, individual words, to analyze the practices. And for the mobile apps we are looking at the source code. So we are not actually running the apps, but rather we are downloading the apps from the Play Store, decompiling them, and look at the source code. And then we compare the results of the two.

I mentioned the word privacy requirements. And privacy requirements are something we came up with. So these are self-defined, and derived from laws. The reason why we are not comparing our results directly to the law is that there are laws that are not applicable to every app. For example, as you all know there are special laws for children, for financial institutions, and this allows us to define ourselves a set of requirements that we want to analyze without necessarily getting into the difficult question of whether an app actually violates the law.

And on the right side of the slide, you see some of the privacy requirements that we analyze. First of all, we require that an app has a privacy policy. And then there are various notice requirements. For example, the notice of policy changes. That is something that we took from the California law. So users have to be notified in case of material policy changes. How they are informed of these changes. And the notices, are something that the privacy policies themselves have to comply with. On the right side you see collection and sharing practices. And those have to be talked about in the privacy policy, as well as implemented in the app. So that is something that applies to both of the policies, as well as the apps.

The first finding that was surprising to me that we have is that many apps don't have a privacy policy. Although they should probably have one. And about half of the apps that we analyzed, did not have a policy. We had a total of 17,991 apps, and out of these, 71 percent did not have a policy, despite processing PII. We used, for the policy analysis, machine learning methods. And for the analysis, static code analysis. I don't want to go into the details here. But I'd be at the post session later, so if you're interested in the details, please stop by, and I go into that.

I just want to talk a little bit briefly about the results that we received here. And the first point I want to make is that the inconsistencies between apps and privacy policies are quite numerous actually. So if, for example, you look at the first row, CID means the collection of device IDs. So

that means that a first party, an app developer, uses an API, to get a device ID. For example, an IP address, or the actual device ID, from an Android phone. You can see that 50% of apps are actually doing that without stating so in their privacy policy, or omitting to write anything about device identifiers. And that is true for all the practices that we looked at. Maybe the sharing of contact information, which you see in the last row is the exception here, but for all the other ones we certainly have higher numbers than we initially expected.

The second point to note here is, again, looking at the first row of collection of device identifiers is that we are able to find all the problems, which you can see from the recall value in the third column, which is one, but we have some false positives. So that means we identify apps that are actually covered by the privacy policy, or the app analysis goes wrong in those cases. And that is what this number of 0.75 in the precision column, of the first row means. And I think that is something we have to improve. But the good news is that by manual work this can be still helpful. And it's probably better that way, to not miss anything, and have some false positives. As opposed to missing problems.

So this is just to give you an idea of the results that we have. And these results I just mentioned were for individual apps. And what you see here is a graph that relates to a group of apps. So if you are interested in finding apps where you have a high chance of finding inconsistencies between a policy, and an app, then this graph tells you should look at apps that do not have a top developer badge, and that have very few user ratings. So these two things, on the Android store, identify apps that have, more often, problems than apps that have a badge, and that have use

when there is a privacy sensitive request is just infeasible, because of insanely high frequency. But in the same work we found out that people do want to have a more final level control over regulation. Not them just being prompted once, and then let the system make the decision. They want to make the decision more and [INAUDIBLE]

The question is that if we ask too many benign questions it's going to habituate the user for

So the important question is that, can we actually capture how people varied that decision based

the user in the process. So this prompting not only makes sure that the system won't make a mistake. It also helps or trains the classifier in subsequent cases, in future similar cases, the system can make the decision on behalf of the user correctly, without involving the user.

So we still have questions to answer. When systems are making decisions on behalf of the user, there is always this chance the system can make the wrong decision. So the question is how we can increase the transparency of this automated decision making, so that the users can go back, and check whether the decisions are being made correctly, whether they are aligned with their own preferences. If not, how they can fix it. The second one is there is the observation of using passively observable traits. This is very significant in the domains of variables, and [INAUDIBLE]. The user environment is very, very restrictive, or impossible. but we still need to learn that preferences. So we can use these passively observable traits in those domains, to learn their preferences, without actually confronting them on every single use case.

While most of the permission models, or the access regulations are moving towards being more restrictive, but as a community we don't have a clear strategy how we can deny access. Are you go2(a)4(r)3(ds)-1(be)4(1)-6(h),uo4(rp)-2(e)4( )-2(e)4(d)-2((y)20( )c4(ut)-2(oo)-40(f)3(i)3(t)4(n))4(npp)-2(e)-2

]TJ -0..Ty39.tl

JUSTIN BROOKMAN: Feel free to jump in if anyone else wants to join, but I have a follow up to the gentleman's second question about observing personal information. And this was the thing we talked about on the planning call, which I thought was interesting, which is about the challenges of encryption, because we like encryption. We recommend encryption as something to safeguard traffic from outside attackers. But in some ways, researchers are sometimes the attackers, right? In, kind of, both of your presentations. And something that we encountered at OTEC, when we looked at smart TVs, we could see the smart TV was phoning home, something, but it's actually sometimes really challenging to do man in the middle, especially on an operating system people don't know very well. So maybe talk a little bit about what some of the challenges are, you guys have seen, as far as encryption. And then whether it does interfere a lot with the research you've been doing. And then, kind of maybe, what the right balance is to kind of make sure that these black boxes, right, are accountable. We can kind of find out what they are saying about us. But still we also like the safeguarding from other people's attacks.

DAVID CHOFFNES: Yeah. I can speak to that. When we started this study in 2015. A lot of things were in plain text. I think over the past year, increasingly, we see information flows transitioning into encryption. And that's good, in terms of the man in the middle eavesdropper. But it is true that, as researchers, we have to go to more and more extreme measures, to be able to understand what is happening inside that traffic. Sorry, inside those encrypted connections. And so both myself, and a number of my colleagues. I'll let Narseo speak for himself. But increasingly we're thinking about ways that we could address this problem through maybe changing how we treat different parts of the data flows. So for example, if a device is leaking information about me, do I have a right to see if that information is leaking about me? And if that's the case, there are technical solutions that would allow you to encrypt it in a way, that the owner of that data, and only the owner of that data, would be able to see that.

But there's definitely going to be some challenges in moving to this environment, because often, when information is sent over the network, some of it may be about a user. Some of it may belong to the app, or to the company, it may be considered sensitive, and they don't want to expose that to researchers or others. So in terms of actually making it happen, I think there would probably need to be a push in terms of policy. But from a technology perspective, we certainly have the basics, and the elements in place to achieve something like this.

NARSEO VALLINA-RODRIGUEZ I second all of what Dave said. But in our experience, we are seeing around 70% of the apps using TLS. And only a handful of them cannot be intercepted. So just with a basic, man in the middle attack, we can decrypt them easily. And I think that the advantage in our site is that many of those applications want to run on corporate environments, where their ideas are deployed. So they have to allow a third party to inject a certificate, and somehow perform man in the middle attacks. So we should take advantage of that for a while.

And, in any case, there are other cases in which you can take more extreme measures. Like, if your Samsung TV is talking to a domain that you're not trusting completely, then you can completely block, and act as a flow firewall, to some extent.

JUSTIN BROOKMAN: The gentleman, at the fifth mic.

SPEAKER 2: Yeah. I have a question for Primal. When you did your field study, did you find that users generally answered for each of the prompt for the permission the same way? And is it possible to maybe crowdsource some of the decision making?

PRIMAL WIJESEKERA: I think that, I feel that [INAUDIBLE] they've already been looked into how to crowdsource these privacy decisions. What we found out with based on our field study is that if they are using-- right now, we don't know what are the full spectrum of contextual circumstances they use to make those decisions. So the question remains, like if we crowdsource, can we account all these circumstances. So when they deny