FTC PrivacyCon 2018
February 28, 2018
Segment 2
Transcript

KRISTEN ANDERSON: Everyone,

First of all, the data aggregation-- a different party may interested in users PII. Then when you send these PII over an encrypted channel it's susceptible to the eavesdropping attack. Now, we've created the definition for privacy. How do we measure that? And still to look at how it evolves over time.

First we need a large data set. We collected 512 Android apps that covered over 7,000 unique versions all across the last eight years. And in order to measure privacy we first need to interact with those apps. It's good if we can do it manually, like a regular user do. However, it's not a very scalable for 7,000 apps. So what we did instead is to combine automated and scripted interactions. We use Monkey to generate a render event, and we went for the case data, we needed to log into the account to use most of functionality. We manually log into the app and replay those events across versions.

So after the interaction, we were able to induce privacy leaks to the internet. And we use manual intermediate proxy to intercept both encrypted and unencrypted traffic. From there, how do we detect a PII? So here we're use the system called a ReCon data [INAUDIBLE] device presented last year, which uses virtual learning algorithm to detect the PII leaks without knowing them ahead of time. And then we manually validated results.

So after this experiment we were able to collect the different parts of the attributes of all the each app. First we have the set of PII types data that leaked by the app. And each app have a different version that is corresponding to their release time that can be ordered chronologically. And for each PII lead, we distinguish whether it's done through encrypted channel versus unencrypted ones. Also it matters whether the first party gather the information, or third party gather this information.

So this is our different privacy aspects about each app. Now, I'll give a concrete example. The Pinterest, it's a really popular app that has to be used by millions of users. Here we have one PII leak at advertising ID. If you look at the last part the ID was leaked 12 times to different parties through encrypted channel. And we have this piece of information for all the PII types leaked by this app.

So the first thing you might notice is that the password leak. Two versions sent a password to an undisclosed third party. And the way if affected millions of users, we reported this to developers. They fix it within a month. So that's one part.

Another significant increase you will notice is that the types of PII leaks increased in more recent versions. This included gender, location, advertising ID, and so on. So not just the unique types. If you look at a specific PII leak, for example, the Android here, the frequency also changed. It was leaked two times in 2016. Then the intensity increased into over 200 times. And this is only 10 minutes into action. So this is really upstanding find going to location tracking.

So not only that, another thing we looked at is the [INAUDIBLE] used to transmit the PII. The location half the time it was sent in a encrypted channel. That's just one app. The take away is that we see lots of variance in privacy leaks across versions. Now we have over 500 apps. What's the big picture? Our folks aggregated findings in our study.

First is the PII leaks. It can change substantially across versions. Not just a number of unique PII types, but also the frequency of leaking specific PII. Another thing we looked at is the issue [INAUDIBLE] adoption in the mobile space. For the mobile apps it's especially harder because they make harder coded protocols, and it requires extra effort to update those domains. And what do we find is that it takes years for the app to adopt [INAUDIBLE] protocol once the domain is supported.

Another thing they started partly checking. We all know it's pervasive and this is a consistent finding, however, it is also evolving over time. We see that the checking IDs, they are moving towards [INAUDIBLE] IDs, which is good. However, we do see evidence of over 100 domains that have the capability to build a permanent linkage between a unique ID they used to track you, and the basic identity related to PII, like email address, real name, gender, location, and so on.

So that's a different aspect. Does this answer the simple question, is privacy getting better or worse? Well, this really depends on how your [INAUDIBLE] privacy and what's important to you. [INAUDIBLE] paper we went into content different aspects, and I highly encourage you to go to check out. Here I only show the case that the combined [INAUDIBLE] from, the PII types, and the destination domains that received those PII. So the curve here, it goes up. The higher value means more risk, which means the combined risks actually wasn't over time.

Again, this is mainly due to more PII were leaked and more domains were contacted. So to conclude, by our definition of a privacy, we find that it [INAUDIBLE] time. However, even if this situation, we still recommended you to update the apps for security reasons. There's a clear need to continuously monitoring your mobile applications using existing data system like ReCon [INAUDIBLE] other research tools.

To this end we also develop a web interface-- a web interface that will provide a customized preferences to help understand how the privacy changed across versions. And this is also important both for the users and us and the developers. Because sometimes they are not necessary for malicious purpose, and they could fix the box based on the findings. So with that, I conclude my talk. Thanks for your attention.

[APPLAUSE]

KRISTOPHER MICINSKI: Is there a little slide advancer thing somewhere? It's this thing, perfect. All right, so I'd like to start out by saying that this was work with many of my collaborators at Maryland. So to introduce this space I want to pose this example app to you. So we have this little app that's going to come up, and it's going to do something like help you order coffee. All right, and the app is going to have a few different buttons that it shows you. And one of those buttons is going to help you locate some coffee shops nearby, something like that. Maybe share the app with your friends.

But then another button is going to help you do some voice ordering or something like that. And I think most of us can agree that if we click the Voice Order Coffee button, then we would probably reasonably expect that our microphone might be used. But maybe we would not expect

But then some could be used more. So there are some permissions where the app just fetches the data before it ends up getting used. And then the fact that it was used, gets shown to users later for example. And then there are some that are really rarely used in an interactive way at all. Things like the power, information about the application, and unique ID of the application. And I'm not really sure why this is, honestly. Although I suspect it might be because it's hard to explain to developers the way that the apps are structured is not amenable to being able to explain them in an intuitive way. But that would be something to look into more.

And next to go along with this app study, I performed a user study. And the thing that we want to measure here is do the user expectations about when these various permissions will be used-- does that align with the patterns that we observed in these applications? So for example, if something is seen very interactively used in the applications that we study, do users also expect that that's the case?

And to do this, users watched slideshows of a whole bunch of different configurations of the way that these apps might use information. So consider that we wanted to answer this question, for example. This was one of our hypotheses, among many others. But we might ask, is a background use of a permission expected after you have had a prior foreground use of that permission?

So you might see a scenario-- or the user would see a scenario-- like this. Where they would first see an app description, and then they would see what we call a user action. And a user action is some event that happens in the application. So in this case, it's that they see the Start screen of the application, and we show them that the Voice Order Coffee button was clicked on. And then immediately they see a screen pops up that asks them for permission to access the microphone, which is what you would see now with an application.

Now we also measured scenarios in which this dialogue didn't appear, to understand if for example, the dialogue was truly necessary under that circumstance. And then we asked them questions about what resources they expect might be used on a Likert scale. From for example, do you think that your microphone is being used right now? Definitely yes or definitely not. And then things like, your location, and then a few other things.

We asked them some distractor questions as well. So that it wasn't just about security. And then we also, after this first event, had a second user action, or a second event that occurred. So in this case the user takes the app and puts it in the background and they see the home screen. And then we ask them those same questions again, maybe in a slightly different order for example. But we might want to measure now, does the user still expect that the microphone might be used when they're on the screen.

Now what we found, the first I think won't bheyNng w(i)-2 (c)4 h(i)-2 (s)-1 (t)-2 (ha)4 (t)-2 (ha)-6 (f)3 (t)21 (e)
omehingd(i)-2 (r)3 (e)-6 (c)4 htlyhi(t)-12 (e)4 (r)3 (a)4 (c)4 (t)-2 (w)2 it it that(c)4 ((e)4 (s)-1 ( )210 (y)20 ( )-10
bwe62 (ha62 (littl(e62 (bw)2 it mow)2 r)52 (e62 (likw)2 (e62 l)-10 (y)-2 (ow)2 ((e62 xw)-8 pw)2 (e62 ce62 t t
ed in the bac ound,dif yu(d(i)-2 don)-10 (')10 (t)-2 ( )-10 ngey tifiatio(f)3 (or)3 (i)-2 (t)-2 . Wce also, ahishisae t

that this policy of asking permission for something the first time it's used can condition users into believing that it might be associated with that given UI element.

So for example, imagine an application that has a map screen. If you ask for permission to use the location on the screen that, for example tracks your run on the map screen. And then you stop your run. You put your phone back in your pocket. Users will be less likely to expect that the location gets used when you're not doing that thing. And I thought that was pretty interesting.

So taking away from this, what can we say? So I think that we can say that for many things that users expect, for example the camera and things like that, applications are already using many of these resources interactively. And we would just recommend that this be the case. So for
2 (i)-2 (c)4 (a)4 (t)-2 (i)-2 (on t)-0 Tw -22.49 -1. for-0 (a)4 (t)-2 2 (v)-10 (e)4 l10 (a)4-2 (hi)--(t)-2 (fb (i)-2 (on c

Ruxpin we had our very first smart toy. He seemed smart. He could read you stories and his mouth moved along in time. And then later we had Furby in the '90s, who learned to speak English. But these toys were not originally connected to the internet. And they certainly weren't the smart toys that we're talking about when we look at these guys today, which are connected to

nice easy link to Facebook and Twitter there. And all of that informed the questions and the responses we eventually got.

And this is Cognitoys Dino again, one of the things that was really good was his mouth lights up and you actually press his belly, which the kid's got a big kick out of, to talk to him. And so for the parents we asked what they expected to be in the toys privacy policy, how they felt about the ability to monitor what their child said to the toy, and would they share what the child said on social media. And this was among many questions, but these were some of the interesting responses we got.

And then we also asked the kids question. What would you talk about with the toy? Do you think the toy can remember what you say to it? Would you tell the toy a secret? And these questions are trying to get at what the kids privacy perspectives are. We were working with kids who were ages 6 to 10, so we didn't want to influence their perspective by asking how do you feel about your privacy. So we tried to elicit that information through these kinds of questions.

This is the email you get from the system when you get it set up. And I think it's really good in that it focuses on giving permission. So THIS isn't just your normal, generic, I accept the terms and conditions, but it actually has in bold right their, it focuses on permission giving. It also explains how they use the recordings and why they need them-- in order to share them with the parents and to improve their services and technology. And it reassures that they don't use these to contact or advertise to children.

But the parents had some important concerns. They wanted to know where the recording was going. They were concerned about who can see it or get the information. On the other side, a parent said look, we have all much stuff to go through right now. I don't have time to go through all these. It's one more pile of media, and I'm not even going to look at it. So we have to rethink how we eb10 (t)-2 4 (c)4 (o)-2 (i) 4 (c)opl2 (e)4 (t)-2 (o g) (t)-2h4 (nd)l2 (e)4 (t)-2 (o h-2 (s)-1 (t)-2 (nf)3 (or)

And do you think your parents could hear what you said to the toy? One kid said, probably. The rest of that quote is, "if it's recording me." And we are like, yep, you got it. You figured this stuff out.

So there were an important connection made by parents and even some of the kids to other

information. So we need to know what factors are important for making this decision. And the third goal is to predict people's privacy preferences so that we can automate privacy decision making. For example, If it can accurately predict that people prefer to turn off an IOT device in a particular situation, our system can just do it for them.

To help us meet these goals we ran an online user study. In our study we asked participants to imagine themselves in some hypothetical data collection scenarios. This kind of a study is often called a vignette study. In a vignette study participants are shown short descriptions of hypothetical situations and are asked some questions to elicit their attitudes toward those situations.

This is an example of a scenario a participant could see. The words inside the brackets are the factors. You're at work. This building has cameras that are recording video of the entire building. The video is shared with law enforcement to improve public safety, and they will not delete it. We instantiated the text in red in different ways. For example, by changing "work" to "home" or "library," or by changing "improve public safety" to "determine possible escape routes" or "optimize the number of staff."

We recorded participants for a 15 minute survey using the Mechanical Turk platform. Mechanical Turk is a crowdsourcing service where people are paid to perform short online tasks, like fill out surveys. We recorded 1007 Mechanical Turk participants from the United States. Each participant was shown 14 scenarios. And after each scenario we asked some questions to understand how often they would want to get notification about data collection in a scenario like this, how comfortable they are with that data collection, and if they would want to allow that data collection.

To interpret their result

beneficial to them. However, they are less likely to want to allow when their data is being shared. Some of the results may seem intuitive to you. In the next few slides I will explain. But just knowing any of these results is not enough to explain people's preferences.

So one question that some of you may have now is, after all these results, what are the most important factors in explaining people's privacy preferences? Is that the data type, the location of

people to help with-- say a survey, or you want somebody to design a logo, you can ask people to help you with that.

And there are various platforms th

But arguably, I mean they don't really need that information to participate in the study. And in this particular case, the participant was very upset about giving up this very sensitive information. And one thing to note is that prior literature suggests that a lot of these are crowdworkers-- they are trying to make ends meet by doing these small tasks, earning some money, oftentimes under minimum wage. So in a way, they're really vulnerable, which makes these sort of issues particularly troubling.

Another type of violation is information processing. As you can imagine, this happens after the data has been collected. So this a very interesting example. The task was about posting pictures of you doing things like smoking and other things. They didn't know, or they didn't tell the workers, they're going to use these pictures for art exhibit. So a month later, they found out and it was a negative surprise.

Another category is information dissemination. This has to do with, once the information has been collected from these tasks, they're going to further disseminate-- either sell them or share with third parties. And the Amazon user policy actually clearly says, "you as a task requester should not request personally identifiable information." and they explicitly say that you cannot collect emails and phone numbers. And low and behold, requesters still do these things. And they still request email. So maybe Amazon should beef up their enforcement of their own policies.

Next category-- invasion. This, in our context refers to things like getting spam messages or even physical stalking. This was one participant talk about over time through different tasks of the same requester, she's giving up both her pictures as well as her address. And which makes her very concerning, because once you have all this information, you compile them together, you can really track individual users.

The last category is what we call deceptive practices. This has to do with a user encounter, phishing attacks, malware, where download some sort of malicious software. So in this particular case this participant was talking about getting phishing attacks. Essentially the task was trying to trick people, disclosing their login credentials.

OK, in addition to the reports we heard from our participants, a member of our research team actually worked as a crowdworker on Amazon Mechanic Turk. And just to get a sense of the kinds of things you might encounter as a real worker. So I'm going to show you some examples. So this is one example where-- I don't know whether you can see those clearly-- basically as the crowdworkers to dig out some personal information for a particular individual. We call this human flesh search. This is interesting because this is a case where it's not the crowdworker's privacy are being violated. I mean, all of us-- any citizen's privacy could be violated if somebody wants to launch a human flesh search on these crowdsourcing platforms.

Another example-- this is the task where they ask crowdworkers to upload videos of themselves, but in order to do that, they have to download a third party software. The thing is, most crowdworkers-- they're probably not very technically savvy. So they're not really in a good position to determine whether the software is malicious or not. So this is a real security risk.

And my final example is this-- it's not from US, it's a Italian example, in fact. And this task is asked the crowdworkers to transcribe a receipt or some sort of form. So one column shows people's real names. The other column next to it is essentially the Italian social security number. This is another example where other people's privacy might be violated through crowdwork platforms.

Another interesting aspect of this example is the looming GDPR. It's come into effect in just a couple of month, and Amazon really have to deal with the issues, because they're dealing with EU citizens. So I think there are two major takeaways from this early research. One is that the kinds of privacy violations we observe from out study are not just reported from workers from a specific geographic location. Meaning that it's very common across the board, which suggests that these issues are system-wide issues. So that's one.

And second, as I talked about earlier, many of these workers are vulnerable. They are there to earn some quick money. And oftentimes they don't really think about their privacy very carefully, which put them at these kind of risks. In terms of implication for technology and policy, we would advocate that there should be tools that enable or empower crowdworkers to be more mindful about potential privacy risks in these tasks. That's not something they're currently considering, partly because the interface on MTurk doesn't really speak to the risks.

Another aspect of tool building is to help requesters. I would tend to think that most requesters are not malicious, but they also haven't really thought about the privacy implications of their requestreqC(na)k8(s)-1 (3t)-2 (-2 (a)4 (t)-2 (i)-2 (a)k81h6dd [(df)3 (K.o)4 (boui)-2 (n6 (y)20 (r)3 (i)--4 (es)-5 (

what consumers are expecting, and what their preferences might be, what might affect that, and how they behave under certain circumstances. So I also invite you guys to ask questions of each other and to interact as much as you'd like.

But just starting off with a couple of questions directed to each of you individually. I'll start with Jingjing. So you found that consumers mobile privacy is less protected now than eight years ago-- that there's more PII collection, leakage, more third-party sharing, slow HTTPs adoption, more cross device tracking. So given all those findings, what do you think are the most important things that consumers need to know about their use of, and their updating of apps? And what are the most effective ways for them to better inform and protect themselves?

JINGJING REN: Yes, so that's a really good question. The first step of actually protecting your privacy is really understand the apps you are using-- what information they are collecting about you. So that's why in the end of the talk I provided these web tool. Among the popular apps we've looked at, we provide all the transparency into the privacy leakage. And also moving forward you should-- consumers should use the container use system like ReCon, Lumen, other re4 (pps)nd t -10 (g -1.15, ot)-215 Td [(r)-1 (4 (on)-t 0.0 6 (t)-6s0 Tc 0 Tw 10.9 6.74)3 (t)-2we R(n t)-2 (2 (r)T

important. And there's no single answer to that. It really evolves for a user's perspective, for a developer' perspective, they want some kind of trade off between the convenience and privacy.

with someone on Facebook, there's this nice little indicator for example, that says your location will be shared with person x for 60 minutes or something like that.

I think those things are moving in the right direction, but I really don't think we have the answer at this point. Maybe we have some shades of where it's going though.
wundsohat.

ateos . Milkng taa eIaa(ue)4 (t)-2 -1.122Td9[(i)-2EMC /P <</MCID 1 >>BD5 -26.89-22Td9[Td [1K

wfycw-n (Mk-7(h)r)4(a(-g)1(0 t-()4 (h)-)2 a)s6 (Irn t)-2 (nw)2 (D)23 (gei)-2)4 (d w)2 (n t)-2 r 60 mt wim Mnnes

IEa tY (a)-6 (CI) (dRI) (dE)1 (Y)-2 N LD S

w wiari di he rih(mg)1 (, t)-2.()-1.16 (w)(i)-2EMC /P <</MCID 1 >>BD7 -26.89-18.83 Td [1S rawe,o.

And I also think having parents talk about privacy expectations with their kids. I've heard some great discussions. Parents have anecdotally told me how they talk about privacy with their kids. Not just hey, I want you to know I could have access to this information, but also you should know that these kinds of things can be recorded. These kinds of things can be collected. So be aware of that when you're using devices, toys, the internet, that kind of thing.

KRISTEN ANDERSON: And I do want to move on to Pardis, but I want to ask a follow-up question of Kris about some of the findings that you made about the indications of recordings. And Kris, what you think about the interaction with the toy where you're pressing on the belly of the Dino or the belt buckle of the Barbie to be able to activate the voice recognition features. What do you think about that interactivity and what it means for notices?

KRISTOPHER MICINSKI: Yeah, I mean, I think that one thing that I would be-- well, maybe this is a discussion we can have after a little bit later, but I would be interested to see-- the kids think that when they're doing this they're interacting with something, but I suspect that if you're an adult, you click on the button or whatever, and it's a picture of your camera. So obviously it's going to take a picture of you. But for children I just don't think that they would have the same mental model. Do you think that children are cued into the-- is it fair to think that they're going to think about their privacy in that way? If I were a kid, I would just want to have fun or

And after notifying them about these data collections, they'll ask some questions to understand your privacy preferences in that specific situation. For example, when they enter a coffee shop we will ask how they would feel if that coffee shop was using a video camera with a face recognition system to identify its frequent customers.

KRISTEN ANDERSON: OK. And we have a question from the audience for you too. "So when looking at privacy preferences for data collection, did you account for different populations? For example, age of users that may not be as sophisticated as other users?"

PARDIS EMAMI-NAEINI: So we had demographic information as one of the factor in our models-- in all of our models. But we did not find any statistically significant result for that demographic information-- like age or where the are from or different education levels.

KRISTEN ANDERSON: OK. Getting Yang in on the fun here. Yang, you found some common privacy concerns among the crowdworkers. And one of your e (h)4m((h 6(m)-2 ()6 (ll o) (ong)3.9 (popul)am)-

KRISTEN ANDERSON: Great. I do want to bring some more questions in from the audience. So I've got one question for the panel. And this is, "a major problem for both users and developers is notice fatigue. How do you suggest developers go about communicating critical information without overload?" I think we've touched upon this topic before, but I want to open it up to all of you to have a deeper discussion.

YANG WANG: So I guess I'll jump in first. I think as many of the research we just heard suggest that it's important to know what kind of information that people would really be concerned about. Because otherwise, if you provide these sort of things, they're going to get overwhelmed. I mean, the developers are very busy. And oftentimes privacy or security might not be where their primary attention is. So helping them understand, under what scenarios what kind of information would it be very concerning for the users will help them.

EMILY MCREYNOLDS: And I think we're facing a new challenge where-- or at least new over the last few years-- where we have these devices that you're talking to that are talking back to you, which is the research we've been doing. And obviously a 10-page privacy policy doesn't work in that situation. You aren't going to have it read out to you.

But I've heard some really interesting anecdotal ways of dealing with this. So if you ask a privacy expert, they'll say, oh, yeah when I ask one of these devices, "what is your privacy policy," it should tell me the privacy policy. Or it should refer me to that. Fine, great, go to the website. You can find our privacy policy here.

But there's really interesting data on the back end they can use instead. I heard of one company that, what they did was when told, hey, you should have a response to, "what is your privacy policy." And they came back a few months later and they had not done that. And so the person

YANG WANG: So if I might add-- I think one thing that the developers should be able to do is, considering user privacy as part of their development process. So it's not and add-on after the fact. So the example I give for Amazon Mechanical Turk, if you embed the privacy information at the time when the requester are making the request, that makes it easier. So imagine you can embed some sort of a privacy feedback to the developers right in there development environment. I think that would make it easy for them to adopt these suggestions.

KRISTEN ANDERSON: Great.

JINGJING REN: Yeah, I think the [INAUDIBLE] interaction is with the app developers, especially the credentials case. They also don't aware of that, because they are included as third party libraries. And they accidentally send out of the credential to these third party. So it's, to their end, it's also their incentive to know the privacy leaks. But I think there's a common challenges. They do not have a good tool to audit to even their own app. So I guess, even for the research committee, and then we should work with developers to build these type of tools. Developers are also users. So we should make it easy for them to understand the privacy implications of the apps they develop.

KRISTEN ANDERSON: We've got a couple of questions. One of them came in from Twitter and another from in the room. And I'm going to combine them into one question, and throw it up

KRISTOPHER MICINSKI: No. I mean, I think this is really hard. But I have a very pessimistic take on this, which is just that there aren't economic implications for leaking privacy from IoT devices and apps. And--

KRISTEN ANDERSON: I think there are implications. Do you mean this is hard to measure?

KRISTOPHER MICINSKI: I'm skeptical. I mean, I always see people's stock bump down, but then it bumps right up back after a while. I mean, I think that if we want to act-- I mean, I think there probably should be economic implications. But I just get the feeling that the vast majority of the time, notions like privacy are so fuzzily defined that it's not-- for obvious egregious things, like leaking somebody's social security number-- sure, that's really bad. But I think the problem is for very nuanced things like maybe leaking some of your location data-- sending it to an analytics library. I just don't think that people view that as that big a deal.

And I think there definitely needs to be more understanding about what users are-- how that economically benefits the app developer. And I think there needs to be this recognition that frequently, app developers and IoT device developers-- they're disincentivized from being transparent with users. And in some sense understanding more privacy stuff can really help us in that way.

But I think it's easy, as many of us have found on the panel, to mask genuine, good things like real benefits of technology. For example, being able to use a toy. Being able to use an app. Being able to get money from a crowdsourcing platform. I think it's easy for that economic benefit to unfortunately let some of the privacy leakage-- let you justify it to yourself that you're going to do it without thinking about the broader implications of maybe losing your privacy.

EMILY MCREYNOLDS: [illegible]

dd.Tc 0 Tw 5 0 Td

a(y)x2 a(hi)ryd