FTC PrivacyCon
January 14, 2016
Segment 5
Transcript

AARON ALVA: All right. Our last panel of the day will look at issues around security and usability as it relates to privacy. So I would like to welcome our first presenter, Sarthak Grover. He is a PhD student at Princeton.

SARTHAK GROVER: Thanks, Aaron. Hi everyone. I'm Sarthak, and I'll be presenting our work on the internet of unpatched things. So the main aim over here is to basically look at the current state of devices. We basically ended up studying network traffic from a bunch of smart devices which are really popular. And we want to talk about how these devices may potentially leak user information. My aim is to basically encourage you to think of how

So what we're interested in right now is what kind of information these common devices leak to the network. And the first device I pick up is the digital photo frame by Pixar. So what we found out was that all traffic from this photo frame is sent in clear text. There's absolutely no encryption happening, all right? The cool thing is that this device can actually talk to your Facebook or RSS feeds. So it's downloading photographs in the clear.

And also, whatever action you take on this device-- for example, you press a button; say you press the play radio button-- that'll actually go in a clear HTTP packet, which somebody, again, on the network can read. So if there's somebody sitting outside on the network, like somebody in the ISP or a malicious passive listener, they can see what you're doing through the photo frame. Apart from that, it's also capable of downloading radio streams-- again, in the clear.

So an example of what kind of information we see-- so these are snapshots from Wireshark, basically. And what we saw was that your email which you configured your account with is actually being sent in clear text. What this means is that this photo frame is potentially leaking account data, and anybody on the network path can actually have a look at this email.

Secondly, if you press the button on this photo frame-- say you press the List Contacts button or the Radio button-- anybody, again, on the network path can have a look at what you currently press. So somebody on the ISP can go, this person is currently listening to the radio from his digital photo frame, though I don't know why you would listen to the radio from the photo frame anyways.

[LAUGHTER]

So basically what I mean to say is that you can find out about the user activity, as well as some account information, just by looking at the network information.

The second device we picked up was a shock security camera. It's a pretty common camera which is used for security monitoring in homes. It has motion detection. What we saw was that all the traffic, again, was being sent in clear text. Now, this security camera actually requires a login. So if you want to view the screen, you're supposed to enter a password. But that doesn't mean the stream itself is encrypted. In fact, anybody sitting on the network can still have a look at where the stream is going and what the stream is. Also, if you go to the web interface and you press a button, whatever you did will still go in an HTTP GET packet, again unencrypted.

So videos are being sent as JPEG frames. Also, if you've pressed the FTP button, then all your data is being uploaded to the FTP, again in the clear. And this is an example of what things look like. So the FTP is actually using some really random ports, so you can't really rely on the network to secure you again, because these are non-standard ports which are being used by the device. This is basically private data which is being uploaded.

The third device we ended up looking at was the Ubi. So this, I think, is like a precursor to the Amazon Echo. Basically, it's a smaller voice box which you can talk to, interface with other devices. For example, we have this Ubi interface with the Nest thermostat in our houses. So what

So this brings us to my conclusion and some implications on the policy. Basically, I don't want to

But it roughly looks something like this. There are three kind of big parties in the picture. So

read other apps' files from external storage. They can try to load them and try to show them to the user, but they cannot actually get access to them directly. They cannot look at their content.

So, so far so good. So it seems like this whole way of protecting users from potentially malicious mobile ads is fairly carefully designed and carefully thought through, except that there is this one little weird thing. They cannot read them, but they can try to load them.

Why is this interesting? It turns out that by trying to load a file that doesn't belong to them, mobile ads can learn a little bit. They can learn like one bit of information. They learn if a particular file exists on the device or not. They cannot read it. They just learn if a file with a particular name exists.

That seems like, OK, all right. That's fascinating. Why am I talking about this? Because that's really a very small amount of information. So now let's look at how this information might be used by a mobile app. So let's take an application which actually has nothing to do with mobile advertising. It's just a popular application in the Google Play Store that happens to be a drug shopping application. So this allows people to go and look at pharmacies. If somebody's picking a particular medication, they can find a pharmacy nearby where the price is lowest on it.

So in this particular case, you can see there are some medications. These particular things-- actually, the fact that a person is taking one of these might be considered sensitive, because this has to do with anxiety and various psychological disorders. So what this app does, if a person is regular shopping for a particular drug, to make it faster it takes a picture of the pill, the literal picture like I'm showing here, and stores that picture in external storage of the device so that next time it's faster to show this picture.

Now imagine that there is an ad running in a different app on the same device. OK? The app that's showing that would be a totally random app. It has nothing to do with the pharmacy shopping app that I showed you before.

However, as I told you before, an ad being shown in it has the ability to ask a very simple question. Does a file with a particular name exist on the external storage? And in this case, it's asking for a file whose name corresponds to the image of one of the anxiety drugs. So what can a mobile ad-- and this is a question to you guys-- learn from the answer to this question? So all it learns is one bit. If the file with a particular name exists on the device, what can the ad learn by knowing the answer?

SPEAKER 2: [INAUDIBLE]

VITALY SHMATIKOV: If the answer to that question is yes, the only reason a file with this name would have existed on this device is the user used that app and searched for that drug. There is no other reason. So if an ad sees that a file like this exists, it cannot read this file. All it needs to know that this file exists it learns with 100% certainty, because this name is unique, that the person has been shopping for a particular drug.

We actually, when we first did this study last summer, we didn't make it public right away

lack industry support and they're not sufficient adoption incentives for companies to actually implement those solutions that have been proposed.

And I'm not going to go too much into the details for the sake of time, but one of the interesting results is that even the experts don't always agree on the interpretation of a privacy policy. And one reason for that is that the policies are vague, but also that they're sometimes contradictory and there are just too many different contexts handled in a single policy.

Good news is that for data collection practices, those are relatively easy to identify and to extract. They're usually in one part of the policy. But data sharing practices are a bit more complicated. They're spread out throughout the policy. Sharing is mentioned in many different contexts and parts of the policies. So it's kind of difficult to extract finer nuances reliably.

Now when we compare the performance of the crowdworkers who are skilled annotators, we actually find quite encouraging results. So when we hold the crowdworkers to a certain quality standard-- 80% agreement, which means eight out of 10 crowdworkers need to come up with the same interpretation-- then we actually find that in a large number of the cases, these crowdworkers agree with the interpretation that our grad students find, as welle ceh7-2(nt)6(t)-6(h)-4(4 Tc 0 acc

that kind of category. And that means that the task interfaces we can show to crowdworkers are a lot more compact, and they can complete those tasks faster and with lower errors.

And based on that, we've developed an annotation scheme that really makes use of this approach. This is an interface not for crowdworkers. We're using this with law students. But the next step is to then break this up again with a project just outlined. But there's a very fine-grained invitation approach, and we're currently collecting data from law students. We already have over 100 policies annotated.

And this provides a really, really rich picture on how information is represented, how data practices are represented, in the policies. We're going to release a data portal to allow exploration of this data on privacy day this year, January 28. So visit our website towards the end of the month. And the nice thing about this data is it's really helpful to train machine learning and natural language processing models, and drive research in this area.

Ultimately what we would be hoping for is that we can actually automate the extraction. And one approach we've been working on here is paragraph sequence alignment. So if I have a paragraph in one policy, in the Amazon policy, and I know that this one's about collection of contact information, and if I compare that paragraph to other paragraphs in other po

And at the same time, we're really interested in understanding what users care about so we can on the one hand, focus the analysis, but also help regulators focus their activities potentially to look at those issues users care about or are concerned with. And at the same time, we want to show ways to effectively inform users about the data practices that are currently lost in those

work on the fixed web, if this already doesn't work on the mobile web, what are the chances that it's going to work in IoT with the Internet of Things. And so our vision in this space, as I said, is this idea that perhaps personalized privacy assistance could be developed that will actually reduce the burden and implore you to manage your privacy better across these different environments.

And so the idea is that these personalized privacy assistance in particular will learn over time your privacy preferences and will be able to semi-automatically configure many of those settings based on various correlations between how you feel about sharing your information with one app versus another app, based on also understanding what your expectations are, going back to the presentation that was given this morning by Ashwini Rao, who's been looking at these issues in particular.

For instance, if you think, as Florian also mentioned, about privacy policies, when you read these privacy policies they tend to be very long, very verbose. But very often, at the end of the day there's only a very tiny fraction of the text in that policy that matters to you, and perhaps even a tinier fraction of the text that pertains to things that you didn't already expect.

And so perhaps this personalized privacy assistance could help us by highlighting those elements of policies that really would be a surprise to us, that perhaps would lead us to modify our behavior as we enter a smart room, for instance, in an IoT context. Perhaps this personalized privacy assistance could also help motivate users to revisit some of their settings, to verify that they still feel the same way. Privacy preferences are not fixed. They might change over time based on experience, based on what you learn.

And so again, what I'd like to do is I'd like to share with you some of our success at actually supporting some early elements of this functionality. What you're seeing here is effectively an early model that we built about how people felt sharing their information with various mobile apps for various steps of purposes, whether the app required this information for internal purposes, for sharing with advertising networks, for profiling purposes, or for sharing with social networks.

I'm not going to describe this chart in great detail because time is limited, but effectively what we're supposed to see here is that people don't always feel the same way on average when it comes to sharing their information. There are clearly di

And so the story here, and the reason why privacy is so complex, is that we don't all feel the same way about these issues. If we did, then it would be simple to come up with defaults and use these defaults for the entire population and it would be done. And perhaps even the FTC could jump in and say, well, nobody feels comfortable about this. Therefore, we're going to outlaw it. Clearly, that's not the way we operate.

And so the reason why this is complex is because we have this diversity in preferences. Some people are quite fine with their fine location being shared with advertisers, and others object. The good news, however-- and this is a result that has come out of our research over the past years-- is that very often, it is possible to organize the population and their preferences into fairly small groups of people that feel very much the same way about these issues.

And so what I want to share with you here is, again, an early example of our work in this area, where again, looking at these mobile app permission preferences we're able to organize a population of users in just four groups. And just based on these four groups and what we're able to predict based on the preferences within each one of these four groups, we're able to show that it might be possible to predict somewhere between 75% and 85% of their privacy preferences when it came to configuring their permission settings.

And so this is very, very simple technology. I'm going to show you that we've been able to go much farther than that. But that gives you a sense already for how easy it is, actually, to predict many different settings that perhaps people would want to have.

So this next chart here shows you the next step in our research in this area, where we looked at actually a population of 240,000 users. I should actually say a population of 3 million users, but we had to clean up the data quite a bit. And we've actually zoomed in on the fraction of the population that was most engaged with their permission settings. So these were LB users LB is a variation of the Android operating system. It was an early version of Android where users could actually configure many different settings.

And we were able to show that through profiles, but also through personalized learning, we could, just by asking people a very small number of questions, effectively predict most of the settings that they would need to configure on their smartphones for the apps that they were going to download. So for instance, if you were to ask them just six questions you could effectively reach a level of accuracy of about 92%. If you're willing to double the number of questions you're asking, you're getting close to 95%.

Now, we are not suggesting in any way that you should fully automate the setting of privacy permissions. We strongly believe in dialogs with users. But there are situations where it's extremely clear how the user feels about some settings. And there are situations where you can determine that actually, your model is not good enough to predict what those settings should be. And that's where you should ask the user. And so that's effectively what we're advocating.

And so we've gone one step further this past summer, and we actually piloted this technology with real users on their actual cell phones. And so we develop profiles-- in this case, I came up with seven different profiles-- and ask people to download this very early version of the

personalized privacy assistant. This assistant would ask them between three and five questions based on the actual apps they had on their cell phones. And based on their answers, it would recommend a number of different settings, as you can potentially see on the right hand side of the slide in front of you.

And so the idea is that the owners of these resources should be able to very simply declare where these resources are deployed and what information these resources collect, and all the other sorts of attributes that you would ideally want to see in a

Or you could imagine a more ambitious effort, where you might say, well, after all, there are actually some interesting correlations between the way you feel about your settings on mobile apps when it comes to sharing information with mobile apps, and perhaps your settings on Facebook, and perhaps your settings in your browser. And so rather than asking you these five or 10 questions in each one of these environments in order to determine what your privacy preferences are, how about just asking these questions perhaps just once, then using a personalized privacy assistant that cuts across all these different environments, interacts with these open APIs to effectively configure many of these settings on your behalf.

So that's our vision in this space. It's not guaranteed that these APIs will be made open. In fact, today they are not. They're very much part of the strategy that some of these larger entities have when it comes to building their ecosystems. But we would like to effectively build an effort towards perhaps convincing these larger players that they would all benefit from opening up these APIs. And perhaps people will ask me questions later on so I get to say more about this, but I think I've run out of time. So thank you very much.

[APPLAUSE]

AARON ALVA: So we'll conclude today with our final discussion of the day. So unlike previous sessions that have focused mostly on privacy, this session has focused on security and usability research as it relates to privacy. So Sarthak discussed security issues related to IoT devices and how they may affect privacy in the home. Vitaly presented on ad libraries and how the lack of tailored security controls in some contexts could result in disclosure of users' information through shared external storage.

For usability, Florian shared about an entire line of research going on around using machine learning, crowdsourcing, and other methods to make privacy policies more usable and for consumers, for businesses, as well maybe for regulators. Finally, Norman presented new ways to understand and manage users' privacy expectations through personal privacy assistance. So overall, this session has provided some new views into different strands of privacy research to consider.

And with that, all of those will add to the policy conversation here. I want to welcome Geoffrey M0(onhg12(an)-4(t)-6)-2(m)-2(,( )-10(I)9)-4( ) E 0 TxJ 0 T-6(t)-2t(pr)3(e)4u3(e)4u3(e)r-6(S)6(l)-2(or)

One of the things I would say is that it's a little bit unfortunate we don't have more economists and engineers talking to each other. As you might have gathered from the last panel, an economist will tell you that merely identifying a problem isn't a sufficient basis for regulating to solve it, nor does the existence of a possible solution mean that that solution should be mandated. And you really need to identify real harms rather than just inferring them, as James Cooper pointed out earlier. And we need to give some thought to self-help and reputation and competition as solutions before we start to intervene.

Now, it is certainly something in the nature of a conference like this-- and for that matter, the kinds of papers that people are writing, because journals don't publish papers saying there's nothing wrong. They publish papers saying there's a problem, and perhaps suggesting solutions

And what are the incentives for consumers themselves? We spend all our time talking about the incentives of firms and the implications of legal liability on firms, but what about the consumers

APIs, clearly this will never be something that one would be able to mandate. But perhaps efforts can be encouraged by bringing together key stakeholders.

At the end of the day, when privacy is presented the right way and when people are looking at this rationally, everyone can benefit from better privacy, including vendors that are sometimes presented as if they didn't care about privacy. I think that if you look, for instance, at what is happening today in the mobile space, it's very clear that everyone has come to realize that they don't want to be seen as the people who don't care about privacy. And that creates strong incentives for them to rethink the way in which they've been approaching some decisions in that space.

So I think that perhaps the FTC can, on the one hand, continue to do what it's been doing very well, I believe, which is to encourage best practices it has done, for instance, for mobile apps, as it has done more recently when it comes to IoT security. And perhaps also convening meetings and encouraging efforts where people look at opportunities for perhaps developing common standards-- not trying to impose any standards.

And standards are very challenging and very tricky efforts, but at least trying to bring together key stakeholders and getting them to think about where they've got effectively common interests and where they might benefit from perhaps developing some open APIs.

VITALY SHMATIKOV: I think transparency is very important. Better understanding and better disclosure of how information is collected and shared between various players in the picture is crucially important, because what we have in mobile space today is these old permission models. They capture something about security of the devices. They capture virtually nothing about privacy. There is a lot of information collection and sharing and information used between all kinds of parties-- platform operators, ad libraries, ad builders, advertisers-- that simply exist outside the existing permission models that a lot of privacy work focuses on.

So the transparency has to be in concert with the right trust model where people want it to be shown in the way that it's comfortable for them. Otherwise, they adapt and your transparency backfires.

AARON ALVA: Norman, did you want to address the transparency with the--

NORMAN SADEH: I'd like to respond to the last comment. So I think it's clear that privacy is an arms race. I think that-- and I worked together with Florian on the project that he described. But the day that site operators, for instance, start modifying their policy based on our technology because of the success of our technology would be a very good day.

We're not quite there yet. If that day happens, we will actually have the ability to probably identify that. And that might potentially be something that the FTC would be interested in. Whether the FTC would necessarily be able to do very much about it or not, I'm not sufficiently versed into the legal ramifications of that but I suspect that it would have something to say if you can establish effectively a pattern where once you effectively are able to capture some practices that are not necessarily putting these companies in good light, they start modifying the way in which they're p

Should we do it again? If so, should we do it exactly the same way? What should we do differently? We'd be very interested in hearing that from you.

One of the things that I would like to do while I'm at the FTC is to try to better bridge the gap between academic research and policymakers. I think the privacy area is an area where there's a real need to inform policymaking with research. And so as such, I look forward to continuing the discussions that we started here throughout the year. Thank you.

[APPLAUSE]

[MUSIC PLAYING]