Session 5: Security and Usability
The Internet of Unpatched Things

Co-author: Nick Feamster (Princeton University)
The Internet of Unpatched Things

Sarthak Grover and Nick Feamster
Princeton University
Current State of Consumer Smart Devices

Many different manufacturers, small startups, novice programmers

Low capability hardware, not enough for security protocols

Most data goes to an online server on the cloud

Even devices in the same home communicate via the cloud
Unpatched IoT Devices Put Our Privacy at Risk

IoT device network traffic:

-Leaks user information
-Identifies the device being used
-May also identify current user activity and behavior!
Case Study of Some Common Home IoTs

- SmartSense Multi-sensor
- Smartthings Hub
- Belkin WeMo Switch
- Nest Thermostat
- Ubi Smart Speaker
- PixStar Digital Photoframe
- Sharx Security IP Camera
- Laptop Gateway (Passive Monitor)

Connections:
- WiFi
- Z-Wave
- Internet
Digital Photoframe: Traffic Analysis

- All traffic and feeds (RSS) cleartext over HTTP port 80
- All actions sent to server in HTTP GET packet
- Downloads radio streams in cleartext over different ports
- DNS queries: api.pix-star.com, iptime.pix-star.com
Photoframe: Privacy Issues

User email ID is in clear text when syncing account.

Current user activity in clear text in HTTP GET.

DNS queries and HTTP traffic identifies a pix-star photoframe.
IP Camera: Traffic Analysis

All traffic over cleartext HTTP port 80, even though viewing the stream requires login password

Actions are sent as HTTP GET URI strings

Videos are sent as image/jpeg and image/gif in the clear

FTP requests also sent in clear over port 21, and FTP data is sent in clear text over many ports above 30,000

DNS query: www.sharxsecurity.com
IP Camera: Privacy Issues

**Video can be recovered** from FTP data traffic by network eavesdropper

DNS query, HTTP headers, and ports identify a Sharx security camera
Ubi: Traffic Analysis

All voice-to-text traffic sent in clear over port 80

Activities sent in clear, and radio streamed over port 80

Sensor readings are synced with server in the background over port 80

Only communication with google API used HTTPS on port 443 and port 5228 (google talk)

Ubi: Privacy Issues

Although HTTPS is clearly available, Ubi still uses HTTP to communicate to its portal. Eavesdropper can intercept all voice chats and sensor readings to Ubi’s main portal.

Sensor values such as sound, temperature, light, humidity can identify if the user is home and currently active.

Email in the clear can identify the user.

DNS query, HTTP header (UA, Host) clearly identifies Ubi device.
Nest Thermostat: Traffic Analysis

All traffic to nest is HTTPS on port 443 and 9543

Uses TLSv1.2 and TLSv1.0 for all traffic

We found some incoming weather updates containing location information of the home and weather station in the clear. 

Nest has fixed this bug after our report.

DNS query: time.nestlabs.com, frontdoor.nest.com, log-rts01-iad01.devices.nest.net. transport01-rts04-iad01.transport.home.nest.com
Nest: Privacy Issues

Fairly secure device: all outgoing personal traffic, including configuration settings and updates to the server, use HTTPS.

*User zip code bug has been fixed

DNS query as well as the use of the unique port 9543 clearly identifies a Nest device.
Smartthings Hub: Traffic Analysis

All traffic over HTTPS on port 443 using TLS v1.2

No clear text port 80 traffic

Flows to an Amazon AWS instance running smartthings server

3-5 packets update every 10 sec in the background

DNS query: dc.connect.smartthings.com
Smartthings: Privacy Issues

Very secure: No information about IoT devices attached to hub is leaked

Background updates every 10 seconds (over HTTPS) fingerprint the hub

DNS query identifies Smartthings hub, but not individual devices
Conclusion: Be Afraid!

Very difficult to enforce security standards

- Multiple manufacturers
- Low capability devices
- Use of non-standard protocols and ports

Difficult to maintain and patch due to low workforce and/or expertise

- Who is responsible? (ISPs? Consumers? Manufacturers?)
- Who is liable? Who should pay?
Conclusion: Be Afraid!

Very difficult to enforce security standards

- Multiple manufacturers
- Low capability devices
- Use of non-standard protocols and ports

Difficult to maintain and patch due to low workforce and/or expertise

- Who is responsible? (ISPs? Consumers? Manufacturers?)
- Who is liable? Who should pay?

Can we solve this on the network? If so, how?

- How much information about user behavior do devices leak to the network?
- Can we offload device security to the home gateway or the cloud?
Thanks!
Vitaly Shmatikov
Cornell Tech

What Mobile Ads Know About Mobile Users
What Mobile Ads Know About Mobile Users

Vitaly Shmatikov

joint work with
Sooel Son and Daehyeok Kim
1.8 million apps in Google Play Store
source: AppBrain

41% include at least one mobile advertising library
source: AppBrain

Every third ad-supported app includes multiple advertising libraries
source: Shekhar et al. (USENIX Security 2012)
Web browser

Web

Mobile app

Mobile

Ad library

YOUR AD HERE

YOUR AD HERE
Prior research

Our focus
Advertising services
• Large businesses
  – AdMob (Google), Mopub (Twitter), AirPush, many others
• Provide AdSDK libraries to 100,000s of developers
• Millions of $ in revenue
• Reputation at stake

Advertisers
• Lots of fly-by-night operators
• Ads resold via auctions, brokers, exchanges
• No reputation at stake, no accountability
• Dynamic filtering and sanitization are hard

Ad libraries must protect users from malicious advertising
Mobile ad impressions are sandboxed inside WebView
Standard Web same origin policy: JavaScript in a mobile ad cannot read or write content from other origins.

... can load (but not read!) files from external storage.
Android External Storage

- Can be read or written by any app with appropriate permissions
- Media-rich mobile ads require access to external storage to cache images, video
- Very weak access control for external storage
  - Any app can read any other app’s files
  - But mobile ads are not apps. **Same origin policy = untrusted JavaScript cannot read ext-storage files**
  - ... but can attempt to load them
1-bit “local resource oracle”: does a file with a given name exist in the device’s external storage?
App for finding pharmacies, compare drug prices (1 to 5 million installs in Google Play Store)

Bookmark functionality

Thumbnail images of drugs that the user searched for cached in external storage
Does this file exist?
file://sdcard/Android/data/com.goodrx/cache/uil-images/45704837

Any ad displayed in any other app on the same device can infer which drugs the user is taking.
This app does **not** include advertising...

... but ads shown in any app on the same device can use the presence of its cached files to infer user’s secrets

Does **not** violate same origin policy
Dolphin mobile browser
(50 to 100 million installs in Google Play Store)

To reduce bandwidth usage and response time, caches fetched images, HTML, and JavaScript in external storage
Any ad displayed in any other app on the same device can infer which sites user visited recently.
Our Study

• Several major Android advertising libraries
  
admob  AdMarvel  
mopub  airpush

• “Local resource oracle” present in all of them
• All acknowledged the issue,
  several fixed in their latest AdSDK releases
How does information about the user flow from AdSDK to advertisers?
Flow of User’s Location in MoPub

1. Register a creative and a tracking URL.

2. Send an ad HTTP request to ads.mopub.com?

3. Deliver the creative that contains the tracking URL with actual values.

4. WebView sends location and id via the tracking request.

- Advertiser
- Advertising creative
- MoPub Servers
- Mobile user
If app has ACCESS_FINE_LOCATION, ad can infer the user’s trajectory
Our Results

• First study of how Android advertising services protect users from malicious advertising

• **Standard Web same origin policy is no longer secure in the mobile context**
  – Mere existence of a certain file in external storage can reveal sensitive information about the user
  – Other security and privacy issues

• Proposed a defense; direct impact on the design of the mobile advertising software stack

See our NDSS 2016 paper
Florian Schaub
Carnegie Mellon University


Co-authors: Norman Sadeh, Alessandro Acquisti, Travis D. Breaux, Lorrie Faith Cranor, Noah A. Smith, Fei Liu, Shomir Wilson, James T. Graves, Pedro Giovanni Leon, Rohan Ramanath, Ashwini Rao (Carnegie Mellon University); Aleecia M. McDonald (Stanford University); Joel Reidenberg, N. Cameron Russell (Fordham University)
Towards Usable Privacy Policies

Semi-automatically Extracting Data Practices from Privacy Policies

Florian Schaub
Carnegie Mellon University

Norman Sadeh | Lead Principal Investigator
Carnegie Mellon University

www.usableprivacy.org

A NSF SaTC Frontier project (CNS-1330596)
Privacy notice & choice

user → privacy policy ← service provider
Privacy notice & choice

Privacy policies

• long & complex
• difficult to understand
• jargon & vagueness
• lack of choices

Privacy policies

- long & complex
- difficult to understand
- jargon & vagueness
- lack of choices

“Only in some fantasy world do users actually read these notices and understand their implications before clicking to indicate their consent”
Privacy policies

• long & complex
• difficult to understand
• jargon & vagueness
• lack of choices

“No one reads privacy policies!

“Only in some fantasy world do users actually read these notices and understand their implications before clicking to indicate their consent”
Overcoming the status quo

- Layered privacy notices
- Privacy nutrition labels
- Privacy icons
- Machine-readable policies (e.g. P3P or Do Not Track)
- ...
Overcoming the status quo

- Layered privacy notices
- Privacy nutrition labels
- Privacy icons
- Machine-readable policies (e.g. P3P or Do Not Track)
- ...

Lack of industry support & adoption incentives
Project objectives

• **Semi-automatically analyze** natural language privacy policies to extract key data practices

• Combine **crowdsourcing, machine learning natural language processing** to enable large-scale analysis of privacy policies

• **Model users’ privacy preferences** to focus on those practices they care about

• **Develop effective user interfaces** that convey relevant and actionable information to users

Tightly interconnected threads

Natural Language Privacy Policies of Websites

Semi-Automated Extraction of Privacy Policy Features

Simplified Privacy Policy Models

Policy Analysis

Inform Public Policy

User Privacy Preference Modeling

Key Features of Privacy Policies

User Privacy Profiles

Effective User Interfaces for Privacy Notices

Inform Internet Users

identification and generation

support personalization

iterative design

Identifying data practices of interest

• **Legal analysis**
  • Analysis of privacy harms addressed through litigation

• **User modeling**
  • Studies on privacy preferences & concerns

• **Policy content analysis**
  • Analysis of how practices are described in privacy policies
  • Ambiguity and vagueness in privacy policies

Crowdsourcing policy annotations

Crowdsourcing policy annotations

collection of contact information

2x Yes: The policy explicitly states that the website might collect contact information

6x Unclear: The policy does not explicitly state whether the website might collect contact information or not
How good are crowdworkers?

• **Studies to compare performance of**
  • privacy policy experts
  • grad students in law & public policy
  • MTurk crowdworkers

• **Annotation of 26 policies**
  • 26 policies annotated by crowdworkers & skilled annotators
  • 6 policies also annotated by experts


How good are crowdworkers?

• Results highlights
  • Even experts do not always agree
  • Data collection relatively easy to identify
  • Data sharing practices more difficult
  • Finer nuances difficult to extract

Accuracy of crowdworker annotations

Compared to skilled annotators on 26 policies

Accuracy levels and agreement counts:
- ≥60% Agree: 151/163 agreement with skilled annotators
- ≥70% Agree: 132/151 agreement with skilled annotators
- ≥80% Agree: 118/123 agreement with skilled annotators
- ≥90% Agree: 88/90 agreement with skilled annotators
- 100% Agree: 42/43 agreement with skilled annotators

Enhancing extraction tasks with Machine Learning and NLP

• Accurate crowdsourcing of policy annotations is feasible

• But privacy policies are still long and complex

• Goal: Help crowdworkers read selectively (thus working more rapidly) without loss of accuracy
Predicting & highlighting relevant paragraphs

Logistic regression based relevance models

Highlight X paragraphs most relevant for current question

Crowdworkers can be induced to label privacy policies faster without affecting accuracy.

Multi-step annotation workflow

Simplified but fine-grained tasks

Click here to read the expanded instructions with an example.

**Short Instructions:** Select the action verbs with your mouse cursor and then press one of the following keys to indicate when the verb describes an act to:
- Press 'c' for **collect** - any act by Zynga to collect information from another party, including the user
- Press 'u' for **use** - any act by Zynga or another party to use or modify information for a particular purpose
- Press 't' for **transfer** - any act by Zynga to transfer or share information with another party, including the user
- Press 'r' for **retain** - any act by Zynga to retain, store or delete information

In the following paragraph, any pronouns "We" or "Us" refer to the game company Zynga, and "you" refers to the Zynga user.

**Paragraph:**
We may **collect** or **receive** information from other sources including (i) other Zynga users who choose to **upload** their email contacts; and (ii) third party information providers.
Annotation dataset

creating corpus of >100 privacy policies annotated by law students

gold standard data for ML/NLP research
Annotation dataset

Creating corpus of >100 privacy policies annotated by law students

Gold standard data for ML/NLP research
Towards automated extraction

Paragraph sequence alignment

Providing notice to users

• **Relevant information**
  - highlight practices users care about
  - emphasize unexpected practices
  - usable and intuitive interface

• **Actionable information**
  - show available privacy choices
  - help users find privacy-friendly alternatives
  - enable users to express dislike of practices

• **Development of Privacy Browser Plugin**
  - provide information independent of website
Browser plugin design

- Display limited set of relevant practices
- User-centered iterative design
  - Focus groups
  - Online studies
  - Field studies
- Public release: Summer 2016
Conclusions

• Semi-automatic analysis of privacy policies with crowdsourcing, natural language processing and machine learning

• Enable large-scale analysis of privacy policies

• Modeling users’ privacy preferences to identify unexpected and relevant practices

• Development of effective user interfaces that convey relevant and actionable information to users

Florian Schaub
fschaub@cmu.edu

Norman Sadeh
Lead PI | sadeh@cmu.edu

USABLE PRIVACY.ORG
To Deny, or Not to Deny: A Personalized Privacy Assistant for Mobile App Permissions

Co-authors: Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Yuvraj Agarwal, Alessandro Acquisti (Carnegie Mellon University)
Personalized Privacy Assistants
From Android Apps to the Internet of Things

Norman Sadeh
Professor, School of Computer Science
Co-Director, MSIT Program in Privacy Engineering
Carnegie Mellon University

www.normsadeh.org  ---  sadeh AT cs.cmu.edu
People Care About Privacy...

Percentages of people surprised by an App’s Permission Requests

- Facebook (rank 1): 60% Location, 0% Device ID, 60% Contact List
- Angry Bird (rank 3): 40% Location, 80% Device ID, 80% Contact List
- Pandora (rank 4): 40% Location, 95% Device ID, 80% Contact List
- Brightest Flashlight (rank 78): 95% Location, 95% Device ID, 95% Contact List

...But They Are Feeling Helpless...

- Privacy policies are too long and too complex

If this has failed on the fixed Web, what are the chances it will work on smartphones or in IoT?
Personalized Privacy Assistants

- **Selectively inform** us about privacy practices we may **not be expecting, yet care about**

- **Learn** many of our privacy preferences and **semi-automatically configure** many settings on our behalf

- **Motivate us** to occasionally revisit some of our preferences and decisions

- The assistants should ideally work across any number of environment and be **minimally disruptive**
One Size-Fits-All Defaults Doesn’t Work

Users’ Average Preferences
- White → comfortable
- Red → uncomfortable

Variance among Users
- Darker yellow → larger variance

Data based on 725 users and 837 apps (>21,000 HITs)
Mobile App Privacy Preferences

A small number of privacy profiles can go a long way

<table>
<thead>
<tr>
<th>Category</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fence-Sitters</td>
<td>75.39%</td>
</tr>
<tr>
<td>Advanced Users</td>
<td>80.54%</td>
</tr>
<tr>
<td>Unconcerned</td>
<td>85.29%</td>
</tr>
<tr>
<td>Conservatives</td>
<td>86.50%</td>
</tr>
<tr>
<td>All Profiles</td>
<td>79.37%</td>
</tr>
<tr>
<td>Grand Average</td>
<td>55.82%</td>
</tr>
</tbody>
</table>

“Grand Average”: Results obtained with “one-size-fits-all” profile
Pure Prediction vs. Interactive Model

Learning personalized privacy preference models

If users can label an additional 10% of their permission decisions, the prediction accuracy will climb from 87.8% to 91.8%...and that’s only 6 questions...

At 20% (about 12 questions), accuracy climbs to 94%!

Data from about 240,000 LBE users, 12,000 apps, 14.5M records

Personalized Privacy Assistant for Android Permissions

These **TRAVEL & LOCAL** apps accessed your **LOCATION** 102 TIMES over the past 2 days:

- Maps
- GasBuddy
- San Francisco
- Yelp
- GrubHub
- Waze

In general, are you OK with **TRAVEL & LOCAL** apps accessing your **LOCATION**?

- YES
- NO

Thank you! Based on your answers, we recommend restricting the following 11 app(s):

- Deny 1 app(s) access to Calendar
- Deny 9 app(s) access to Location
  - Facebook (50 times)
  - News & Weather (0 times)
  - Contacts+ (28 times)
  - Messenger (16 times)
  - Snapchat (84 times)
  - QR Code Reader (0 times)
  - Skype (0 times)

Why deny? This Social app accesses your Location for App Functionality and Consumer Tracking & Profiling.

Do you want to make these changes?

- YES, DENY THE 8 APP(S) SELECTED
- NO, DO NOT MAKE ANY CHANGES
Nudging Users for 6 days

Are users just being nice or is this truly reflecting their preferences?
Successfully Piloted with Android Users

- Piloted with 29 Android users – 10 day study
- Users accepted 73.7% of our recommendations
- Only 5.6% of accepted recommendations were modified over the next 6 days, despite nudges to revisit earlier decisions
  - Users showed great engagement, modifying many settings not covered in the recommendations
- Users are comfortable with the recommendations and see the value of the assistants

“To Deny, or Not to Deny: A Personalized Privacy Assistant for Mobile App Permissions,” Bin Liu, Mads Schaarup Andersen, Florian Schaub, Norman Sadeh, Hazim Almuhimedi, Yuvraj Agarwal, Alessandro Acquisti - working paper, 2016
Extending this to IoT
Personalized Privacy Assistants for IoT

- Registries enable owners to register their IoT resources
  - Resources associated with locations/areas
  - Menus lead to automated generation of machine-readable privacy policies

- PPA’s discover relevant resources by consulting registries & compare policies against user profiles (expectations and preferences)
  - Selective alerts & semi-automated configuration of available privacy settings
Concluding Remarks- I

- PPAs aim to provide a **pragmatic approach to notice and choice**
  - Leveraging machine learning and privacy profiles
  - Learning people’s privacy preferences and expectations to **minimize user burden**, yet ensure that **users are informed** about those issues they care about and **retain sufficient control over their settings**
Concluding Remarks - II

- **Assumption**: Privacy profiles and learned preferences should **only be used for the purpose of managing user privacy**

- PPAs have to come with **strong privacy guarantees**
  - Could be offered by entities controlling specific ecosystems
  - Could be offered by 3rd parties dedicated to privacy management
    - Opens the door to PPAs that cut across multiple ecosystems/environments but **requires open APIs**
Contact: sadeh AT cs.cmu.edu

Acknowledgement: Funding provided under DARPA’s Brandeis initiative, NSF SaTC/SBE Program, Google Web of Things Expedition

Contact: sadeh AT cs.cmu.edu

Collaborators: Bin Liu, Jialiu Lin, Mads Scharup Andersen, Florian Schaub, Alessandro Acquisti, Yuvraj Agarwal, Lujo Bauer, Lorrie Cranor
Discussion of Session 5

Discussants:

• **Aaron Alva**, Federal Trade Commission

• **Geoffrey Manne**, International Center for Law and Economics

• **Davi Ottenheimer**, Institute for Applied Network Security

Presenters:

• **Sarthak Grover**, Princeton University

• **Vitaly Shmatikov**, Cornell Tech

• **Florian Schaub**, Carnegie Mellon University

• **Norman Sadeh**, Carnegie Mellon University
Closing Remarks
Lorrie Cranor, Chief Technologist