Carnegie Mellon University

Automated Analysis of Privacy Requirements for Mobile Apps

Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Peter Story, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin, Joel Reidenberg

Tuesday, February 28, 2017

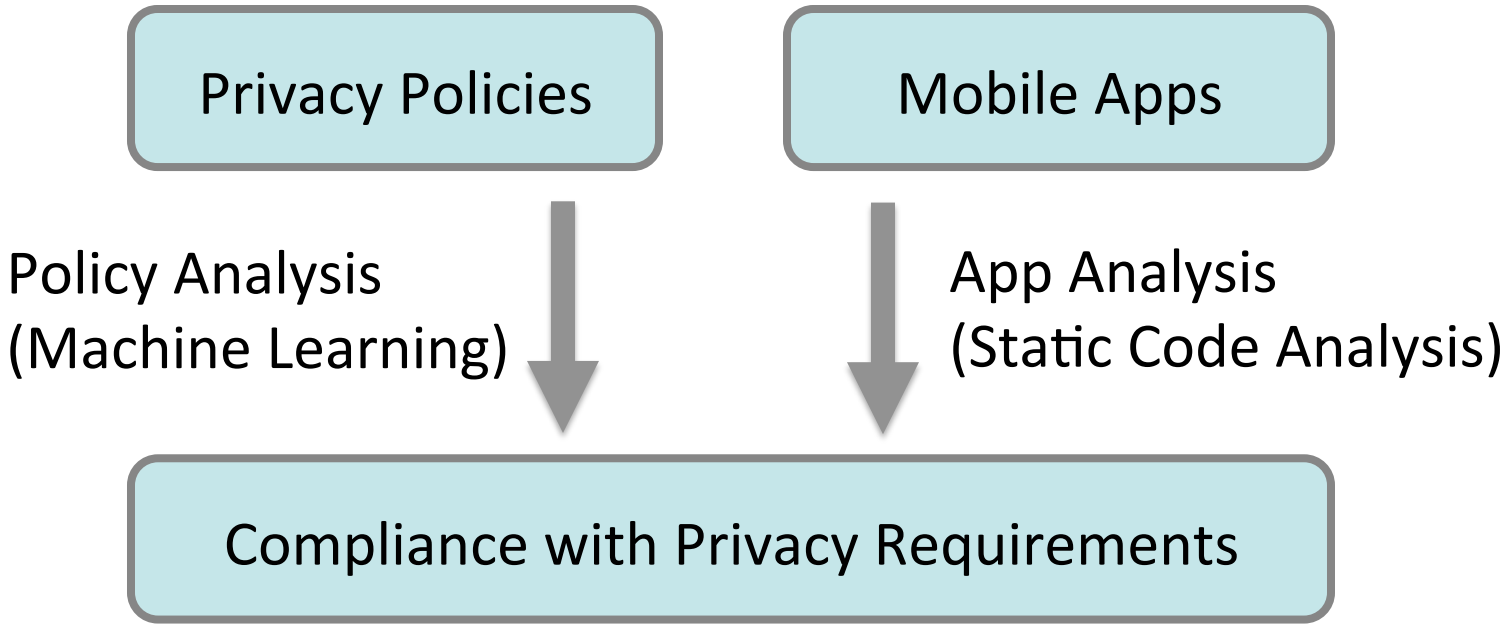
Network and Distributed System Security Symposium 2017
San Diego, California

Motivation

- "Google Play requires developers to provide a valid privacy policy when the app requests or handles sensitive user or device information." (Google Play Developer E-Mail, Feb '17)
- The California Online Privacy Protection Act also requires app publishers to have a privacy policy and transparently disclose data practices (California Business and Professions Code Sections 22575-22579)

→ System to evaluate how many apps have a privacy policy, whether the policies follow privacy requirements, and analyze discrepancies between apps and policies to increase transparency at scale

Compare App Behavior/Code to Policy Text



Analysis Techniques

```
1  def location_feature_extraction(policy):
2
3   data_type_keywords = ['geo', 'gps']
4   action_keywords = ['share', 'partner']
5   relevant_sentences = ''
6   feature_vector = ''
7
8   for sentence in policy:
9    for keyword in data_type_keywords:
10    if (keyword in sentence):
11     relevant_sentences += sentence
12
13  words = tokenize(relevant_sentences)
14  bigrams = ngrams(words,2)
15
16  for bigram in bigrams:
17   for keyword in action_keywords:
18    if (keyword in bigram):
19     feature_vector += bigram, bigram[0],
           bigram[1]
20
21  return feature_vector
```

Binary Classifiers

Policy Analysis

Permission Extraction

Call Graph Creation

Call ID Analysis

APP Analysis

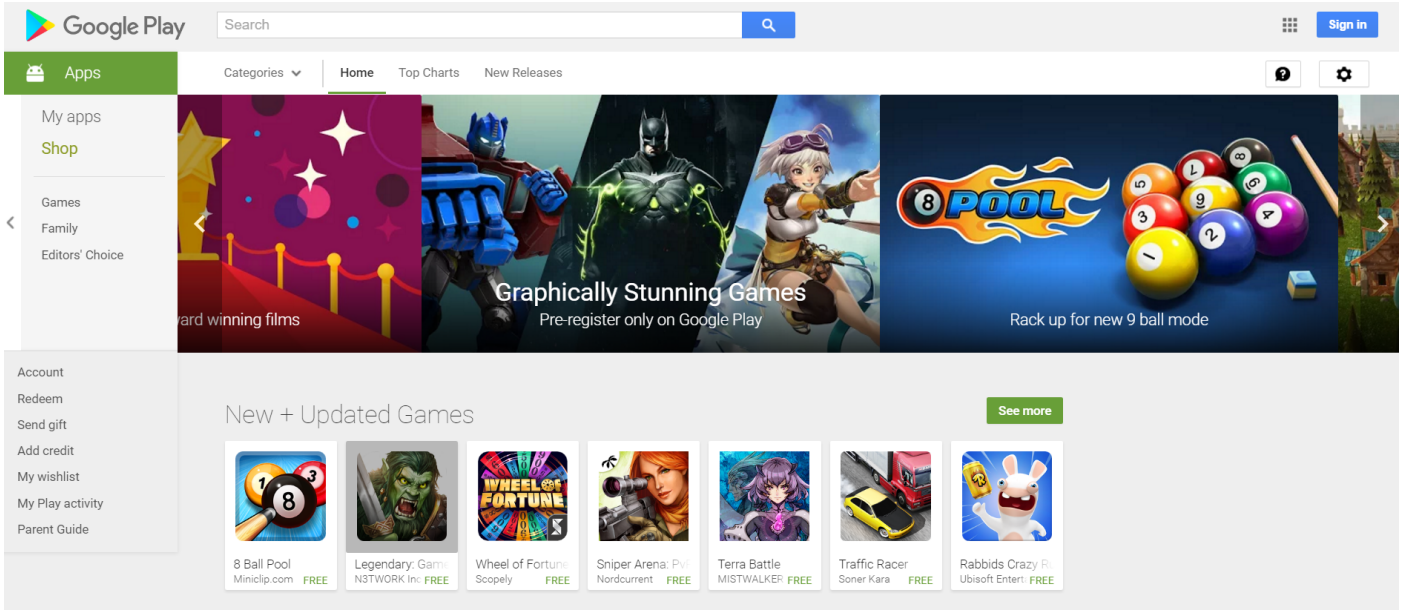
What are privacy requirements?

1. Apps must have a privacy policy

2. Policies have to describe data practices occurring in the apps (e.g., describe how location data is shared with third parties) and must not omit any practice

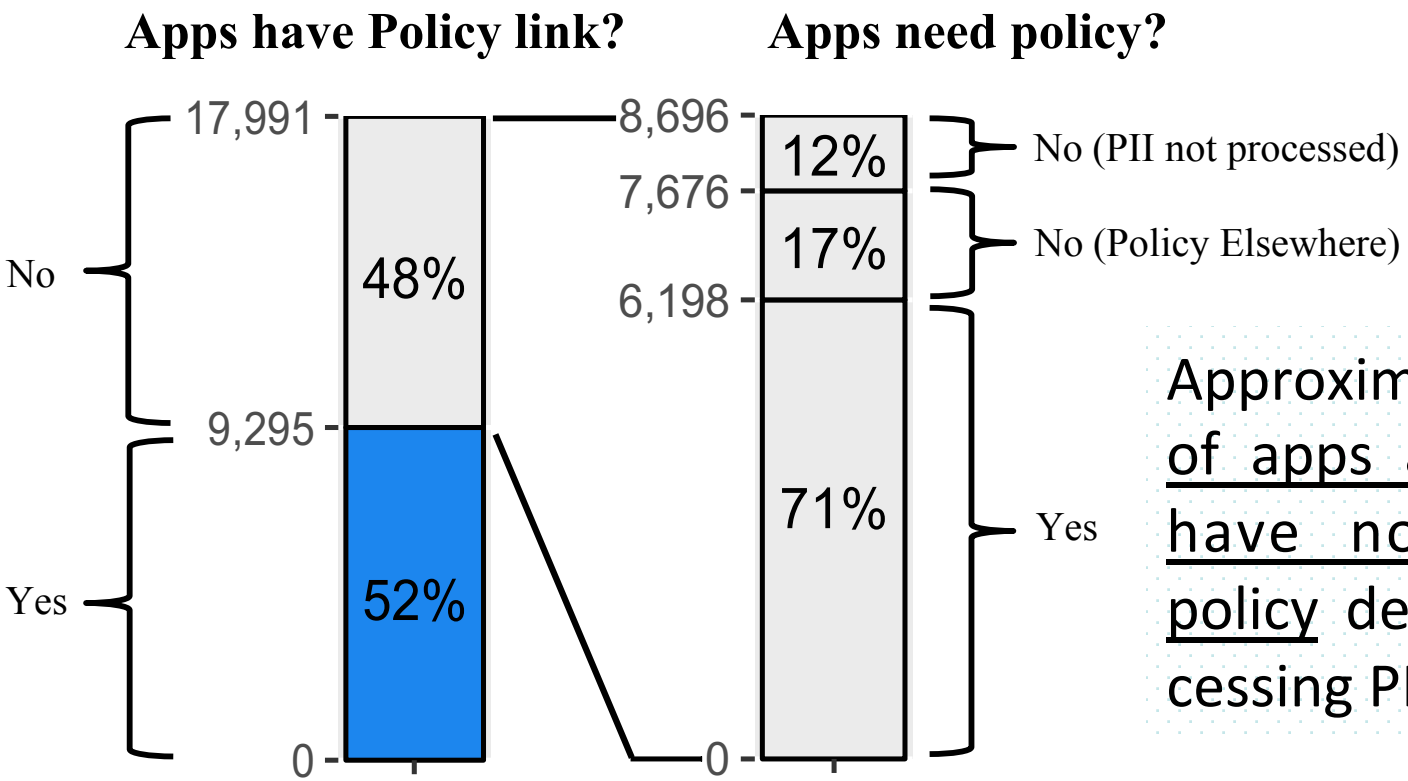
3. Apps must follow the described practices

Dataset



- 17,991 free apps from the Google Play Store and their metadata (e.g., whether an app has a policy link or the number of reviews)
- Started crawl from most popular apps in each category and followed links to similar apps

Potential Privacy Requirement Non-Compliance



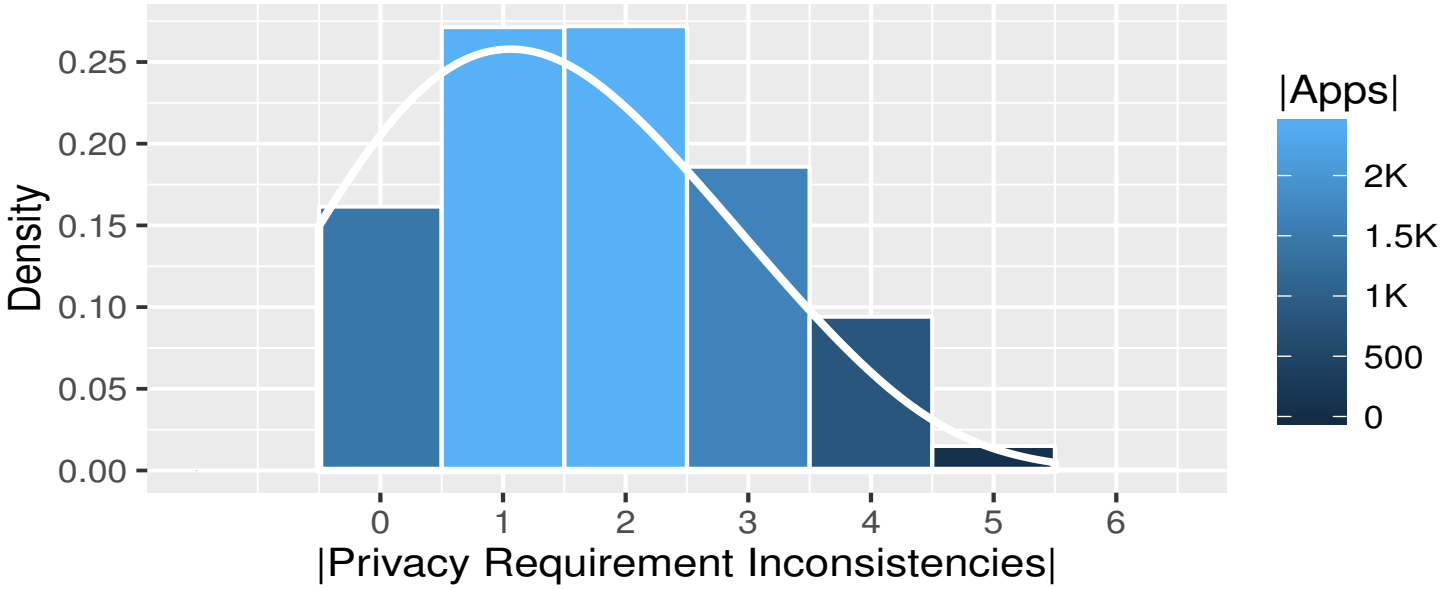
Approximately 1/3 of apps appear to have no privacy policy despite processing PII

Potential Privacy Requirement Non-Compliance

Practice	***Precision*** **(Test Set; *n=40*)**	***Recall*** **(Test Set; *n=40*)**	***F-1*** **(Test Set; *n=40*)**	***% Potential Privacy Requirement Non-compliance (n=9K)***
Notice of Policy Changes	0.96	0.89	0.93	**46%**
Collection of Identifiers	0.75	1	0.86	**50%**
Sharing of Location	1	1	1	**17%**
Sharing of Contact	1	1	1	**2%**

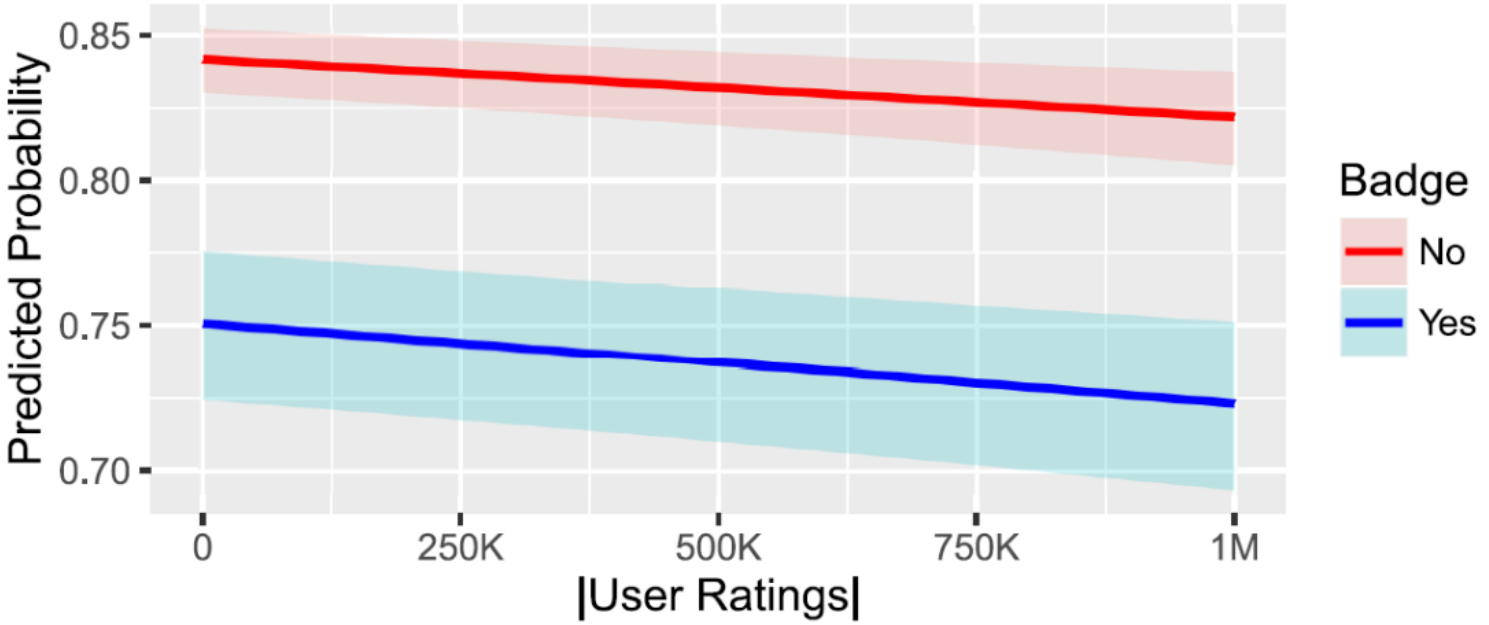
→ Potential privacy requirement non-compliance can be predicted reliably and at scale

Potential Privacy Requirement Non-Compliance



- Each app exhibits a mean of 1.83 instances of potential privacy requirement non-compliance
- Non-compliance does not necessarily mean that a law is violated; manual verification required

Potential Privacy Requirement Non-Compliance



→ Use app metadata to predict which app populations have increased probability of potential privacy requirement non-compliance

Concluding Thoughts

- Help developers, app store owners, and regulators; implement our system into their workflow
- Current system piloted by the Office of the California Attorney General
- Extensions towards other use cases, particularly, in the emerging Internet of Things domain

Thank you!

USABLEPRIVACY.ORG
the usable privacy policy project

National Science Foundation Secure and Trustworthy Cyberspace (SaTC) Project Lead PI: Norman Sadeh (sadeh@cs.cmu.edu)
This material is based upon work supported in part by the National Science Foundation under grants CNS-1330596, CNS-1330214, and SBE-1513957, as well as by DARPA and the Air Force Research Laboratory, under agreement number FA8750-15-2-0277. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, the Air Force Research Laboratory, the National Science Foundation, or the U.S. Government.