Bootstrapping Privacy Compliance in Big Data Systems (cont’d) + Inferring Data Associations in Black-Box Systems

Giulia Fanti
Based on slides by Anupam Datta
CMU

Fall 2019
Administrative

- HW2 will be released this week
  - Stay tuned

- Recitation on Friday (James)
  - More info about project categories
  - Open office hours

- Project proposals due next Friday, Sept. 20
  - Use Piazza to find partners!
Quiz on Canvas

- Take the quiz on your laptops/tablets/devices
- Please do not look back at your notes
- 10 minutes
Last time (continued)

Bootstrapping Privacy Compliance in Big Data Systems

S. Sen, S. Guha, A. Datta, S. Rajamani, J. Tsai, J. M. Wing
Proceedings of 35th IEEE Symposium on Security and Privacy
May 2014.
Formal Semantics

Recursively check exceptions
ALLOW clauses have DENY clauses as exceptions
Top Level clause determines Blacklist/Whitelist
Designed for Precision

\[
\begin{align*}
T^G & \not\subseteq T^C \\
\text{ALLOW } T^C \text{ EXCEPT } D_1 \cdots D_m & \text{ denies } T^G \quad (A_1) \\
T^G & \subseteq T^C \quad \exists_i D_i \text{ denies } T^G \\
\text{ALLOW } T^C \text{ EXCEPT } D_1 \cdots D_m & \text{ denies } T^G \\
T^G & \subseteq T^C \quad \forall_i D_i \text{ allows } T^G \\
\text{ALLOW } T^C \text{ EXCEPT } D_1 \cdots D_m & \text{ allows } T^G \\
\bot & \in T^G \cap T^C \\
\text{DENY } T^C \text{ EXCEPT } A_1 \cdots A_m & \text{ allows } T^G \quad (D_1) \\
\bot & \notin T^G \cap T^C \quad \exists_i A_i \text{ allows } T^G \cap T^C \\
\text{DENY } T^C \text{ EXCEPT } A_1 \cdots A_m & \text{ allows } T^G \quad (D_2) \\
\bot & \notin T^G \cap T^C \quad \forall_i A_i \text{ denies } T^G \cap T^C \\
\text{DENY } T^C \text{ EXCEPT } A_1 \cdots A_m & \text{ denies } T^G \quad (D_3)
\end{align*}
\]

TABLE I
GRAMMAR FOR LEGALEASE

TABLE III
INFEERENCE RULES FOR LEGALEASE
Designed for Expressivity (Bing, October 2013)

ALLOW EXCEPT

DENY $\text{DataType}$ IPAddress:Expired
DENY $\text{DataType}$ UniqueIdentifier:Expired
DENY $\text{DataType}$ SearchQuery, PII InStore Store
DENY $\text{DataType}$ UniqueIdentifier, PII InStore Store

DENY $\text{DataType}$ BBEPData UseForPurpose Advertising

DENY $\text{DataType}$ BBEPData, PII InStore Store

DENY $\text{DataType}$ BBEPData:Expired

DENY $\text{DataType}$ UserProfile, PII InStore Store

DENY $\text{DataType}$ PII UseForPurpose Advertising
DENY $\text{DataType}$ PII InStore AdStore

DENY $\text{DataType}$ SearchQuery UseForPurpose Sharing EXCEPT
ALLOW $\text{DataType}$ SearchQuery:Scrubbed

- “we remove the entirety of the IP address after 6 months”
- “[we remove] cookies and other cross session identifiers, after 18 months”
- “We store search terms (and the cookie IDs associated with search terms) separately from any account information that directly identifies the user, such as name, e-mail address, or phone numbers.”
- “we do not use any of the information collected through the Bing Bar Experience Improvement Program to identify, contact or target advertising to you”
- “we take steps to store [information collected through the Bing Bar Experience Improvement Program] separately from any account information we may have that directly identifies you, such as name, e-mail address, or phone numbers”
- “we delete the information collected through the Bing Bar Experience Program at eighteen months.”
- “we store page views, clicks and search terms used for ad targeting separately from contact information you may have provided or other data that directly identifies you (such as your name, e-mail address, etc.).”
- “our advertising systems do not contain or use any information that can personally and directly identify you (such as your name, email address and phone number).”
- “Before we [share some search query data], we remove all unique identifiers such as IP addresses and cookie IDs from the data.”
ALLOW
EXCEPT
  DENY *DataType PII UseForPurpose Sharing*

EXCEPT
  ALLOW *DataType PII:OptIn*
EXCEPT
  ALLOW *AccessByRole Affiliates*
EXCEPT
  ALLOW *UseForPurpose Legal*

DENY *DataType DoubleClickData, PII*
EXCEPT
  ALLOW *DataType DoubleClickData, PII:OptIn*

"We do not share personal information with companies, organizations and individuals outside of Google unless one of the following circumstances apply:"
"We require opt-in consent for the sharing of any sensitive personal information."
"We provide personal information to our affiliates or other trusted businesses or persons to process it for us."
"We will share personal information [if necessary to] meet any applicable law, regulation, legal process or enforceable governmental request."
"We will not combine DoubleClick cookie information with personally identifiable information unless we have your opt-in consent."
Legalease Usability

Survey taken by 12 policy authors within Microsoft Encode Bing data usage policy after a brief tutorial

Time spent
- 2.4 mins on the tutorial
- 14.3 mins on encoding policy

High overall correctness
A Streamlined Audit Workflow

Legal Team
Crafts Policy

Privacy Champion
Interprets Policy

Legalease
A formal policy specification language

Grok
Data inventory with policy labels

Checker

Developer
Writes Code

Audit Team
Verifies Compliance

Encode

Refine

Annotated Code

Update Grok

Potential violations

Annotated Code

Fix code
A Streamlined Audit Workflow

Legal Team
Crafts Policy

Encode

Privacy Champion
Interprets Policy

Refine

Legalease
A formal policy specification language

Annotated Code

Potential violations

Grok
Data inventory with policy labels

Code analysis, developer annotations

Checker

Developer
Writes Code

Fix code

Audit Team
Verifies Compliance

Update Grok

Legalease Policy
Map-Reduce Programming Systems

Scope, Hive, Dremel
Data in the form of Tables

Code Transforms Columns to Columns
No Shared State
Limited Hidden Flows

users =
SELECT _name, _age FROM datasetAB
user_tag =
SELECT GenerateTag(_name, _age)
FROM users
OUTPUT user_tag TO datasetC
Grok

Purpose Labels
Annotate programs with purpose labels
Grok

Purpose Labels
Annotate programs with purpose labels

Initial Data Labels
Heuristics and Annotations

```
users = 
    SELECT Name, Age FROM datasetAB
user_tag = 
    SELECT GenerateTag(_name, _age) FROM users
OUTPUT user_tag TO datasetC
```
Groks

Purpose Labels
Annotate programs with purpose labels

Initial Data Labels
Heuristics and Annotations

Flow Labels
Source labels propagated via data flow graph

D. E. Denning. “A lattice model of secure information flow”
A Lattice of Policy Labels

- If “Profile” use is allowed then so is everything below it
- If “Name” use is denied then so is everything above it
Implicit flows

Beyond direct flows discussed in healthcare audit examples
Map-Reduce

Map
Operate on rows in parallel
eg. filtering

Reduce
Combine groups of rows eg. aggregation

users =
SELECT Name, Age
FROM datasetAB

users_35 =
SELECT _name, _age
FROM users
WHERE (_age > 35)

ages_35 =
SELECT _age, COUNT(_name) AS Profile
FROM users_35
GROUP BY _age

OUTPUT ages_35 TO datasetC
## Combine Noisy Sources

<table>
<thead>
<tr>
<th>Carefully curated regular expressions</th>
<th>Expensive</th>
<th>Auditor Verification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leverages developer conventions</td>
<td>Low Noise</td>
<td>Definitive</td>
</tr>
<tr>
<td>Significant Noise</td>
<td></td>
<td>Need very few of these</td>
</tr>
<tr>
<td>Variable Name Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Developer Annotations</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

20
Why Bootstrapping Grok Works

Pick the nodes which will label the most of the graph

~200 annotations label 60% of nodes

A small number of annotations is enough to get off the ground.
Scale

- 77,000 jobs run each day
  - By 7000 entities
  - 300 functional groups
- 1.1 million unique lines of code
  - 21% changes on avg, daily
  - 46 million table schemas
  - 32 million files
- Manual audit infeasible
- Information flow analysis takes ~30 mins daily
Nightly Compliance Process

Static code analysis

Generate report

Manual Audit

Privacy elements: 300K+

Audit candidates: 10K+

Teams: 8

Files: 25M+

Schemas: 2M+
A Streamlined Audit Workflow

Legal Team
Crafts Policy

Encode

Privacy Champion
Interprets Policy

Refine

Legalease
A formal policy specification language

Grok
Data inventory with policy labels

Code analysis
Developer annotations

Checker
Annotated Code
Legalease Policy
Update Grok
Potential violations

Developer
Writes Code

Fix code

Audit Team
Verifies Compliance

Fix code
A Streamlined Audit Workflow

**Workflow** for privacy compliance

**Legalease**, usable yet formal policy specification language

**Grok**, bootstrapped data inventory for big data systems

**Scalable** implementation for Bing
Part II: Inferring Data Usage of Black-Box Systems
So far

- Technique for auditing privacy policies automatically
- Given access to:
  - Developers
  - Code
  - Privacy advocates in the company
- This is really for companies to audit **themselves**
  - Maybe law enforcement
What if we don’t have access?

- LaTanya Sweeney
What was hard about this study?

- Manual ad checking
  - Limits scale of the study

- She knew what she was looking for
  - Associations between black-sounding names and ads for arrest records
  - Limits scope of the study
XRay: Enhancing the Web’s Transparency with Differential Correlation

M. Lecuyer, G. Ducoffe, F. Lan, A. Papancea, T. Petsios, R. Spahn, A. Chaintreau, R. Geambasu

Proceedings of 2nd USENIX Security Symposium
August 2014.
Goals

- Fine-grained and accurate data tracking
  - Detect which inputs (e.g., emails) likely triggered which outputs (e.g., ads)

- Scalability
  - E.g., track past month’s emails

- Extensibility, generality, self-tuning
  - Limited manual tuning when you switch to general websites
Forms of Targeting

1) Profile Targeting

2) Contextual Targeting

3) Behavioral Targeting
Figure 2: The XRay Architecture.
Browser Plugin

- Tracks specific DOM elements in audited services’ web pages
- Which elements to track is configuration setting
  - E.g., Gmail
  - Inputs: Emails
  - Outputs: Ads
Shadow Account Manager

- (1) Populate **shadow accounts** with subsets of user account’s tracked inputs
- (2) Periodically retrieves outputs from each audited service for each shadow account

- These are service-specific:
  - E.g. Gmail
    - Send emails with SMTP
    - Call the ad API
Differential Correlation Engine

- Analyzes correlations in the Correlation DB
- Plugin makes a `get_assoc` request
  - Look up entry in Correlation DB, return pre-computed associations
  - If none found, return `unknown`
- Periodic updates
How do we detect a correlation?

- **Naïve solution:**
  - Create shadow account with every possible combination of inputs
  - Q: If I have $N$ initial inputs and $M$ initial outputs, how many shadow accounts do I need?

  **Emails**
  1. Subject: This job is hard
  2. Subject: Request for help
  3. Subject: Call for papers
  
  .
  
  N. Subject: Canvas isn’t working

  **Ads**
  1. Learn 2 code!
  2. Work from home, earn $500 a day
  
  .
  
  M. Amazon

  **A:** $2^N$. We want every possible subset of inputs
Instead: Set Intersection

- Create $C \ln N$ shadow accounts
- Pick probability $\alpha \in (0,1)$
- Randomly place each input into each shadow account w.p. $\alpha$

Emails
- 1. This job is hard
- 2. Request for help
- 3. Call for papers
- N. Canvas isn’t working

Ads
- 1. Learn 2 code!
- 2. Work from home, earn $500 a day
- M. Amazon
Instead: Set Intersection

- **Given output** $O_k$:
  - Compute set $A_k$ of active accounts that saw $O_k$
  - Compute inputs that appears in fraction $\beta$ of active accounts
  - Return set of accounts iff $\geq \beta$ contain all remaining inputs

**Emails (Inputs)**
1. This job is hard
   - N. Canvas isn’t working

**Ads (Outputs)**
1. Learn 2 code!
2. Work from home, earn $500 a day
   - M. Amazon
Why should this work?

- Key idea: argue that every non-targeting input would have a vanishingly small probability of being in a significant fraction of active accounts

- Try to prove this yourself before next class

- Connections to the idea of group testing
  - Technique from WWII for blood testing
Extension

- To get rid of parameter tuning \((C, \alpha, \beta)\), they introduce Bayesian inference-based detection mechanism

- Behavioral Targeting
  - Defines a generative model for observations, computes likelihood
  - Uses same method of data collection as before

- Contextual targeting
  - Compute likelihood based on assumptions about
    - \(p_{in} = P(\text{see ad} \mid \text{targeted input is present})\)
    - \(p_{out} = P(\text{see ad} \mid \text{targeted input is not present})\)
    - \(p_0 = P(\text{see ad} \mid \text{no targeting})\)
  - Iteratively train parameters, then likelihoods

- Composite model
  - Arithmetic mean of scores
Experimental Methods

- Implemented in 3,000 lines of Ruby
  - Google, YouTube, and Amazon
  - Service-specific shadow account manager
    - ~500 lines of code each

- Ground truth exists for ads on Amazon and YouTube
  - “Why recommended”

- Google labelled manually
Results: Self-Targeted Ads (Sanity Check)

- Check for Gmail targeting via AdWords

<table>
<thead>
<tr>
<th>Ad Keyword</th>
<th>Targeted Email</th>
<th>Detected by XRay?</th>
<th>XRay Scores</th>
<th># Accounts &amp; Displays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chaldean Poetry</td>
<td>Like Chaldean Poetry?</td>
<td>Yes</td>
<td>0.99, 1.0</td>
<td>13/13, 1588/1622</td>
</tr>
<tr>
<td>Steampunk</td>
<td>Fan of Steampunk?</td>
<td>Yes</td>
<td>0.99, 1.0</td>
<td>13/13, 888/912</td>
</tr>
<tr>
<td>Cosplay</td>
<td>Discover Cosplay.</td>
<td>Yes</td>
<td>0.99, 1.0</td>
<td>13/13, 440/442</td>
</tr>
<tr>
<td>Falconry</td>
<td>Learn about Falconry.</td>
<td>Yes</td>
<td>0.99, 1.0</td>
<td>13/13, 1569/1608</td>
</tr>
</tbody>
</table>
Bayesian Model Accuracy

- Experiment on Gmail

(a) Recall

(b) Precision
Bayesian vs. Set Intersection Comparison

(a) Recall

(b) Precision
# Results: Examples of Targeted Ads

<table>
<thead>
<tr>
<th>Topic</th>
<th>Targeted Ads</th>
<th>XRay Scores</th>
<th># Accounts &amp; Displays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer</td>
<td>Black Mold Allergy Symptoms? Expert to remove Black Mold.</td>
<td>0.99, 0.05</td>
<td>9/9, 61/198</td>
</tr>
<tr>
<td></td>
<td>Adult Assisted Living. Affordable Assisted Living.</td>
<td>0.99, 0.99</td>
<td>8/8, 12/14</td>
</tr>
<tr>
<td>Cancer</td>
<td>Ford Warriors in Pink. Join The Fight.</td>
<td>0.96, 0.98</td>
<td>9/9, 1022/1106</td>
</tr>
<tr>
<td></td>
<td>Rosen Method Bodywork for physical or emotional pain.</td>
<td>0.98, 0.05</td>
<td>7/7, 24/598</td>
</tr>
<tr>
<td>Depression</td>
<td>Shamanic healing over the phone.</td>
<td>0.99, 0.99</td>
<td>16/16, 117/117</td>
</tr>
<tr>
<td></td>
<td>Text Coach - Get the girl you want and Desire.</td>
<td>0.93, 0.04</td>
<td>7/7, 31/276</td>
</tr>
<tr>
<td>African American</td>
<td>Racial Harassment? Learn your rights now.</td>
<td>0.99, 0.2</td>
<td>10/10, 851/5808</td>
</tr>
<tr>
<td></td>
<td>Racial Harassment, Hearing racial slurs?</td>
<td>0.99, 0.2</td>
<td>10/10, 627/7172</td>
</tr>
<tr>
<td>Homosexuality</td>
<td>SF Gay Pride Hotel. Luxury Waterfront.</td>
<td>0.99, 0.1</td>
<td>9/9, 50/99</td>
</tr>
<tr>
<td></td>
<td>Cedars Hotel Loughborough, 36 Bedrooms, Restaurant, Bar.</td>
<td>0.96, 1.0</td>
<td>8/8, 36/43</td>
</tr>
<tr>
<td>Pregnancy</td>
<td>Find Baby Shower Invitations. Get Up To (60% Off) Here!</td>
<td>0.99, 1.0</td>
<td>9/9, 22/22</td>
</tr>
<tr>
<td></td>
<td>Ralph Lauren Apparel. Official Online Store.</td>
<td>0.99, 0.6</td>
<td>10/10, 85/181</td>
</tr>
<tr>
<td></td>
<td>Clothing Label-USA. Best Custom Woven Labels.</td>
<td>0.99, 1.0</td>
<td>9/9, 14/14</td>
</tr>
<tr>
<td></td>
<td>Ranchos Official Site</td>
<td>0.99, 0.99</td>
<td>9/9, 0/0</td>
</tr>
</tbody>
</table>
Results: Scalability

(a) Scalability with Input Size
(b) Recall with Input Size
(c) Precision with Input Size

Figure 8: Scalability. (a) Number of accounts required to achieve the knee accuracy for varied numbers of inputs. (b), (c) Recall/precision achievable with the number of accounts in (a). Behavioral uses the Bayesian algorithm.
What are some of the challenges?

- Only detect correlation, not causation
- Required manual tuning for each service