Understanding Privacy and Contextual Integrity: A Personal Journey

Anupam Datta

Carnegie Mellon University

PrivaCI Symposium, Princeton University

September 14, 2018

Princeton, NJ 2018

CENTER FOR INFORMATION TECHNOLOGY POLICY
AT PRINCETON UNIVERSITY

DLI

SYMPOSIUM ON APPLICATIONS OF CONTEXTUAL INTEGRITY

September 13-14, Princeton University.

Co-sponsors:

- Center for Information Technology Policy, Princeton University
- Digital Life Initiative, Cornell Tech.

Attendance by invitation-only.

Goals today

- A personal history of thinking about contextual integrity and privacy
- Challenges and opportunities

Stanford, CA 2005

PORTIA *Privacy, Obligations, and Rights in Technologies of Information Assessment*

People

Academic PIs:

- Dan Boneh, Stanford University
- Joan Feigenbaum, Yale University
- Stephanie Forrest, University of New Mexico
- Hector Garcia-Molina, Stanford University
- Ravi Kannan, Yale University (2003-2007)
- John Mitchell, Stanford University
- Rajeev Motwani, Stanford University
- Helen Nissenbaum, New York University
- Avi Silberschatz, Yale University
- Rebecca Wright, Rutgers University

PRIVACY AS CONTEXTUAL INTEGRITY

Helen Nissenbaum*

Abstract: The practices of public surveillance, which include the monitoring of individuals in public through a variety of media (e.g., video, data, online), are among the least understood and controversial challenges to privacy in an age of information technologies. The fragmentary nature of privacy policy in the United States reflects not only the oppositional pulls of diverse vested interests, but also the ambivalence of unsettled intuitions on mundane phenomena such as shopper cards, closed-circuit television, and biometrics. This Article, which extends earlier work on the problem of privacy in public, explains why some of the prominent theoretical approaches to privacy, which were developed over time to meet traditional privacy challenges, yield unsatisfactory conclusions in the case of public surveillance. It posits a new construct, "contextual integrity," as an alternative benchmark for privacy, to capture the nature of challenges posed by information technologies. Contextual integrity ties adequate protection for privacy to norms of specific contexts, demanding that information gathering and dissemination be appropriate to that context and obey the governing norms of distribution within it. Building on the idea of "spheres of justice," developed by political philosopher Michael Walzer, this Article argues that public surveillance violates a right to privacy because it violates contextual integrity; as such, it constitutes injustice and even tyranny.


Coupa
Café
www.coupacafe.com

Privacy and Contextual Integrity: Framework and Applications

Adam Barth Anupam Datta John C. Mitchell
Stanford University
{abarth, danupam, jcm}@cs.stanford.edu

Helen Nissenbaum
New York University
helen.nissenbaum@nyu.edu

2006 IEEE Symposium on Security and Privacy

Descriptive component of contextual integrity

"In a context, the flow of information of a certain type about a subject (acting in a particular capacity/role) from one actor (could be the subject) to another actor (in a particular capacity/role) is governed by a particular transmission principle."

Nissenbaum 2004

Privacy Regulation Example (GLB Act)

Sender role

Subject role

Financial institutions must notify consumers if they share their non-public personal information with non-affiliated companies, *but the notification may occur either before or after the information sharing occurs*

Attribute

Recipient role

Transmission principle

Formalizing contextual informational norms

$$\sigma \models \Box \forall p_1, p_2, q : P. \forall m : M. \forall t : T.$$

$$\text{incontext}(p_1, c) \wedge \text{send}(p_1, p_2, m) \wedge \text{contains}(m, q, t) \rightarrow \bigvee_{\varphi^+ \in \text{norms}^+(c)} \varphi^+ \wedge \bigwedge_{\varphi^- \in \text{norms}^-(c)} \varphi^- \quad (1)$$

positive norm: $\text{inrole}(p_1, \hat{r}_1) \wedge \text{inrole}(p_2, \hat{r}_2) \wedge \text{inrole}(q, \hat{r}) \wedge (t \in \hat{t}) \wedge \theta \wedge \psi$

negative norm: $\text{inrole}(p_1, \hat{r}_1) \wedge \text{inrole}(p_2, \hat{r}_2) \wedge \text{inrole}(q, \hat{r}) \wedge (t \in \hat{t}) \wedge \theta \rightarrow \psi$

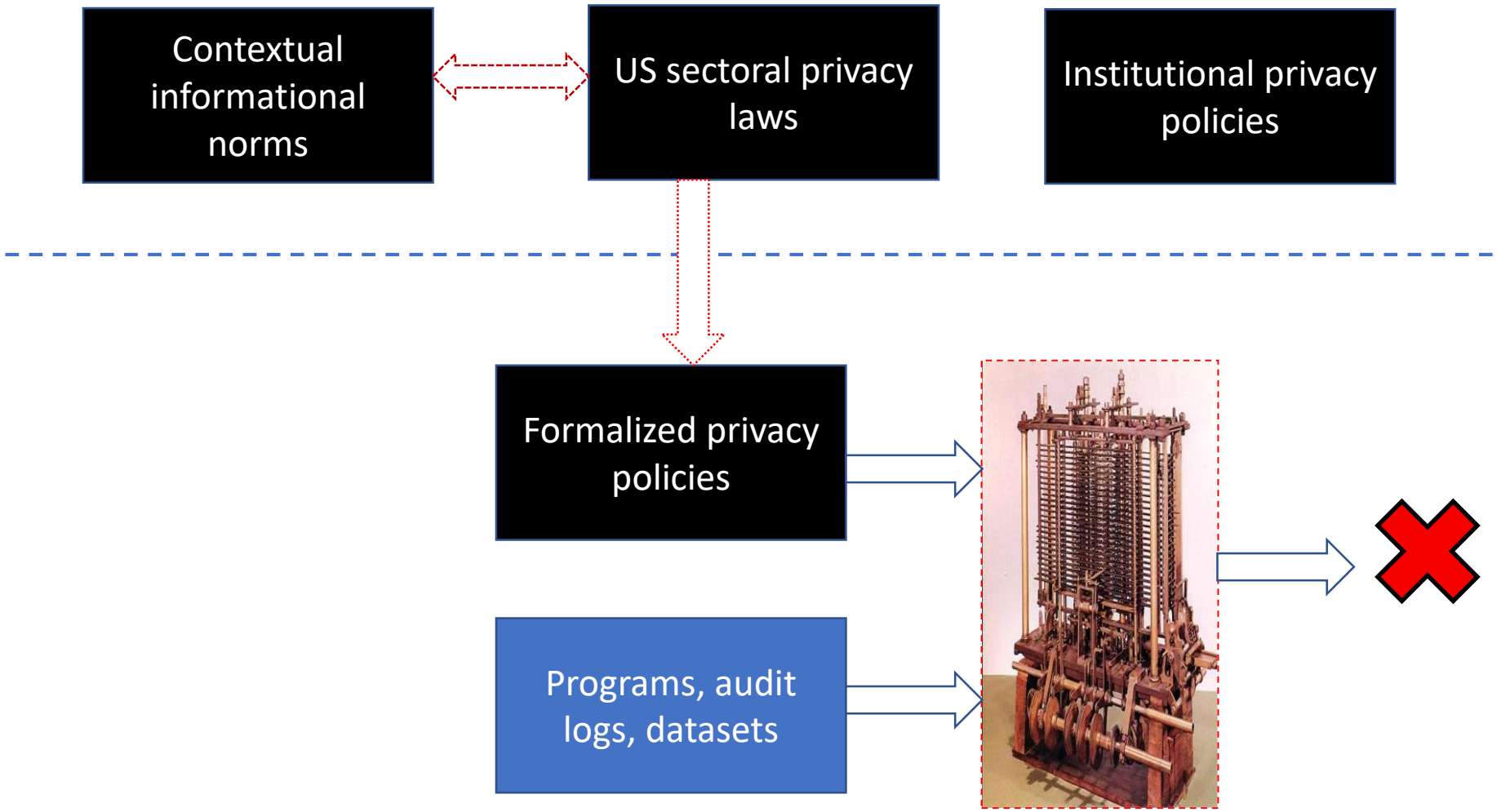
Figure 1. Norms of Transmission Represented as a Temporal Formula

GLBA clause formalized

$$\text{inrole}(p_1, \textit{institution}) \wedge \text{inrole}(p_2, \textit{non-affiliate}) \wedge \text{inrole}(q, \textit{consumer}) \wedge (t \in \textit{npi}) \rightarrow$$

$$\Diamond \text{send}(p_1, q, \textit{privacy-notice}) \vee \diamondminus \text{send}(p_1, q, \textit{privacy-notice})$$

Enforcing privacy



Privacy and Contextual Integrity: Framework and Applications

Adam Barth Anupam Datta John C. Mitchell Helen Nissenbaum
Stanford University New York University
{abarth, danupam, jcm}@cs.stanford.edu helen.nissenbaum@nyu.edu

- Formalized descriptive component of contextual integrity using first-order temporal logic
- Demonstrated that sample clauses from US privacy regulations – HIPAA, GLBA, COPPA – lined up with this form of specification
- Methods for automated monitoring for propositional temporal logic specifications of contextual informational norms

CMU, PA 2007-

Can we specify the entirety of privacy laws like HIPAA and GLBA using this kind of formalism?

Can we (largely) automatically enforce these kinds of privacy policies?

Experiences in the Logical Specification of the HIPAA and GLBA Privacy Laws

Henry DeYoung
hdeyoung@cs.cmu.edu

Deepak Garg
dg@cs.cmu.edu

Limin Jia
liminjia@cmu.edu

Dilsun Kaynar
dilsun@cs.cmu.edu

Anupam Datta
danupam@cmu.edu

Carnegie Mellon University, Pittsburgh, PA 15213 USA

- Complete specification of HIPAA and GLBA privacy laws
- Structure of laws largely follows CI flow descriptions
- Restrictions on use of personal information for specific purposes (beyond CI flow norms)

2010 ACM Workshop on Privacy in an Electronic Society

Example from HIPAA Privacy Rule

A covered entity may disclose an individual's protected health information (phi) to law-enforcement officials for the purpose of identifying an individual if the individual made a statement admitting participating in a violent crime that the covered entity believes may have caused serious physical harm to the victim

- Concepts in privacy policies
  - **Actions:** send(p1, p2, m)
  - **Roles:** inrole(p2, law-enforcement)
  - **Data attributes:** attr_in(prescription, phi)
  - **Temporal constraints:** in-the-past(state(q, m))
  - **Purposes:** purp_in(u, id-criminal))
  - **Beliefs:** believes-crime-caused-serious-harm(p, q, m)

Black-and-white concepts

Grey concepts

Policy Auditing over Incomplete Logs: Theory, Implementation and Applications

Deepak Garg
dg@cs.cmu.edu

Limin Jia
liminjia@cmu.edu

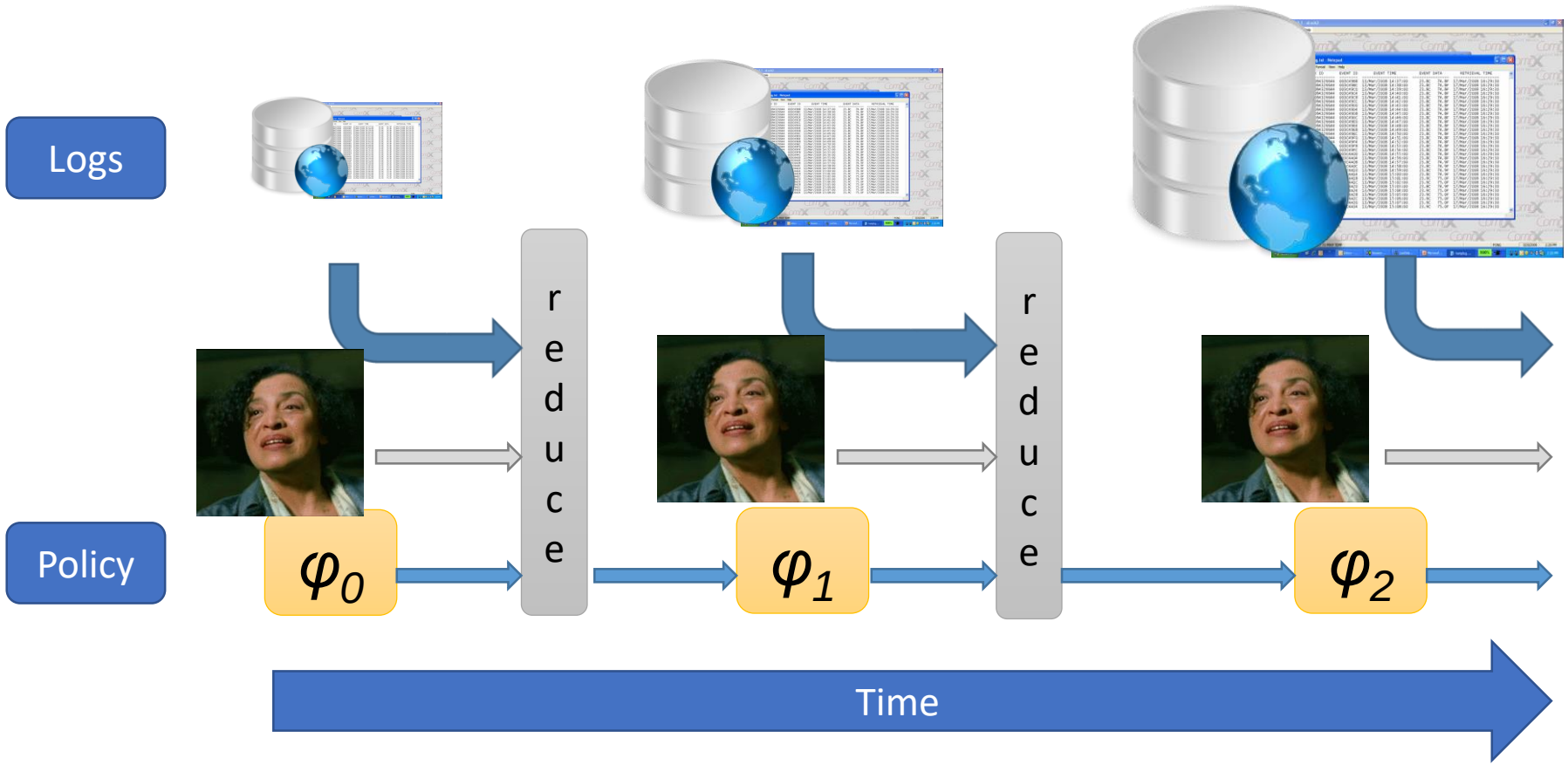
Anupam Datta
danupam@cmu.edu

- Audit algorithm that applies to expressive fragment of first-order logic (cf. propositional LTL in BDMN'06)
- Covers entirety of HIPAA Privacy Rule
- Deals with incompleteness in logs (e.g., subjective predicates about beliefs and purposes)

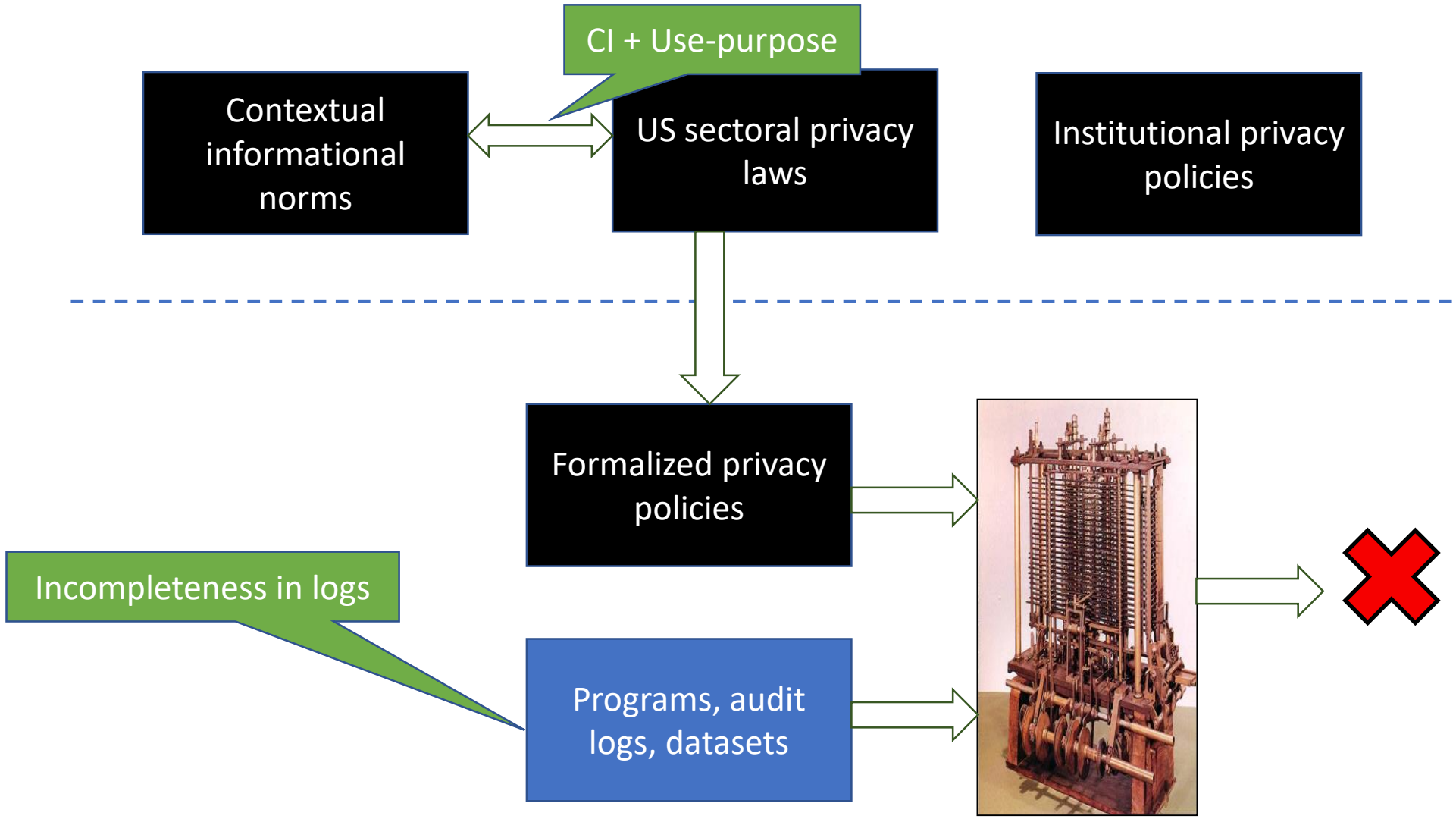
2011 ACM Conference on Computer and Communications Security

reduce: The Iterative Algorithm

$$\text{reduce}\,(\mathcal{L}, \varphi) = \varphi'$$



Enforcing privacy



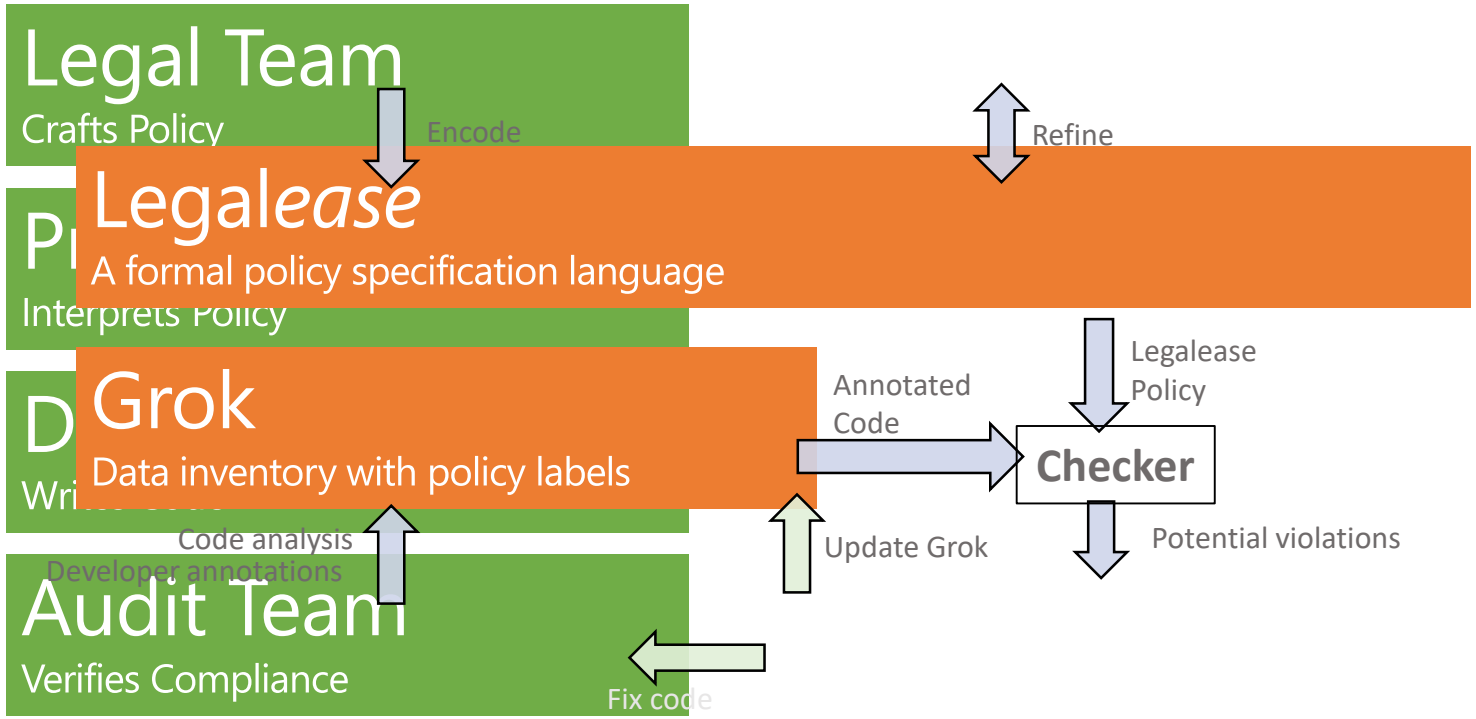
Bootstrapping Privacy Compliance in Big Data Systems

Shayak Sen*, Saikat Guha†, Anupam Datta*, Sriram K. Rajamani†, Janice Tsai‡ and Jeannette M. Wing‡
*Carnegie Mellon University, Pittsburgh, USA
{shayaks,danupam}@cmu.edu
†Microsoft Research, Bangalore, India
{saikat,sriram}@microsoft.com
‡Microsoft Research, Redmond, USA
{jatsai,wing}@microsoft.com

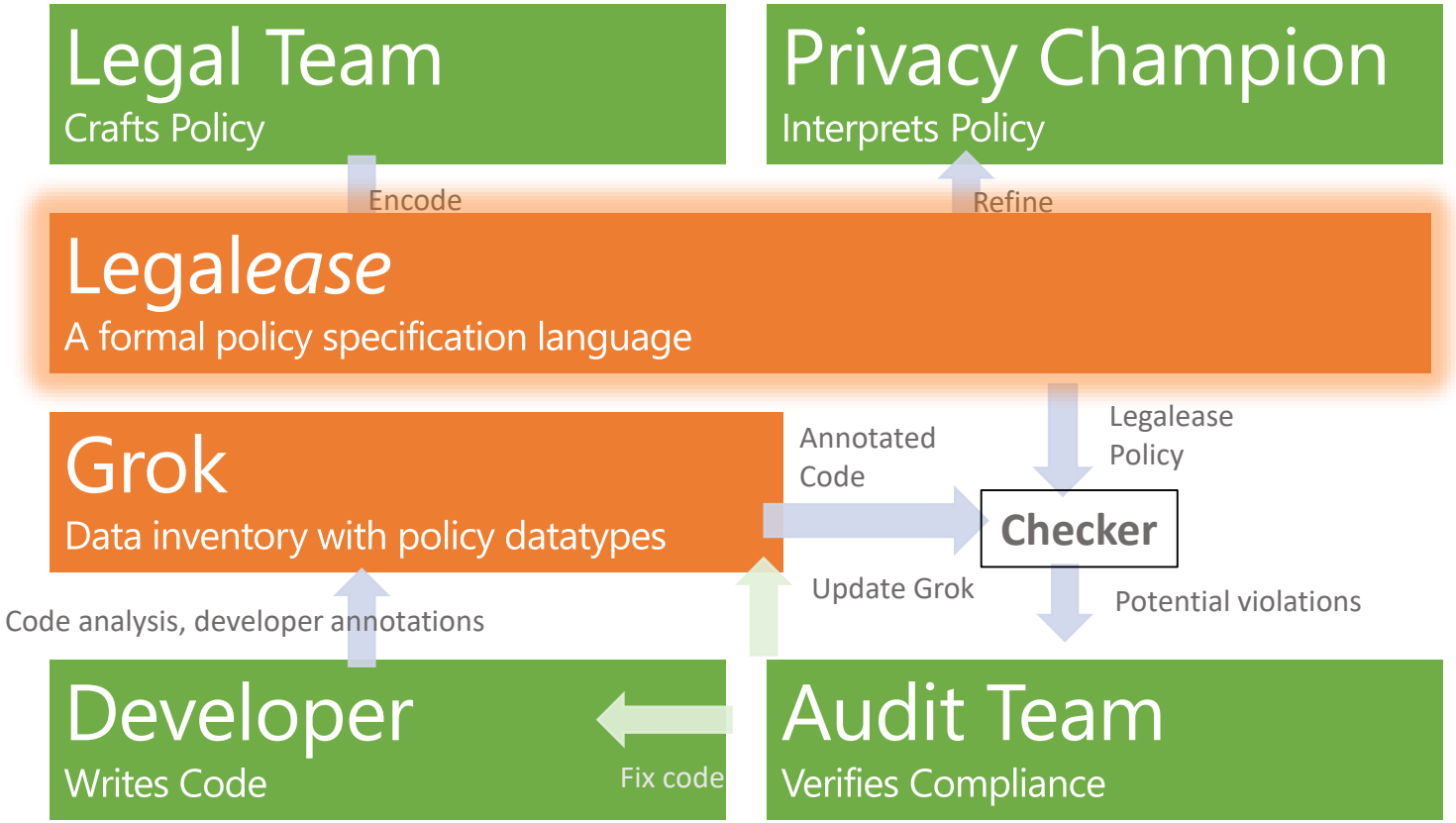
***Abstract*—With the rapid increase in cloud services collecting and using user data to offer personalized experiences, ensuring that these services comply with their privacy policies has become a business imperative for building user trust. However, most compliance efforts in industry today rely on manual review processes and audits designed to safeguard user data, and therefore are resource intensive and lack coverage. In this paper, we present our experience building and operating a system to automate privacy policy compliance checking in Bing. Central to the design of the system are (a) LEGALEASE—a language that allows specification of privacy policies that impose restrictions on how user data is handled; and (b) GROK—a data inventory for Map-Reduce-like big data systems that tracks how user data flows among programs. GROK maps code-level schema elements to datatypes in LEGALEASE, in essence, annotating existing programs with information flow types with minimal human input. Compliance checking is thus reduced to information flow analysis of big data systems. The system, bootstrapped by a small team, checks compliance daily of millions of lines of ever-changing source code written by several thousand developers.**

2014 IEEE Symposium on Security and Privacy

A Streamlined Audit Workflow



A Streamlined Audit Workflow



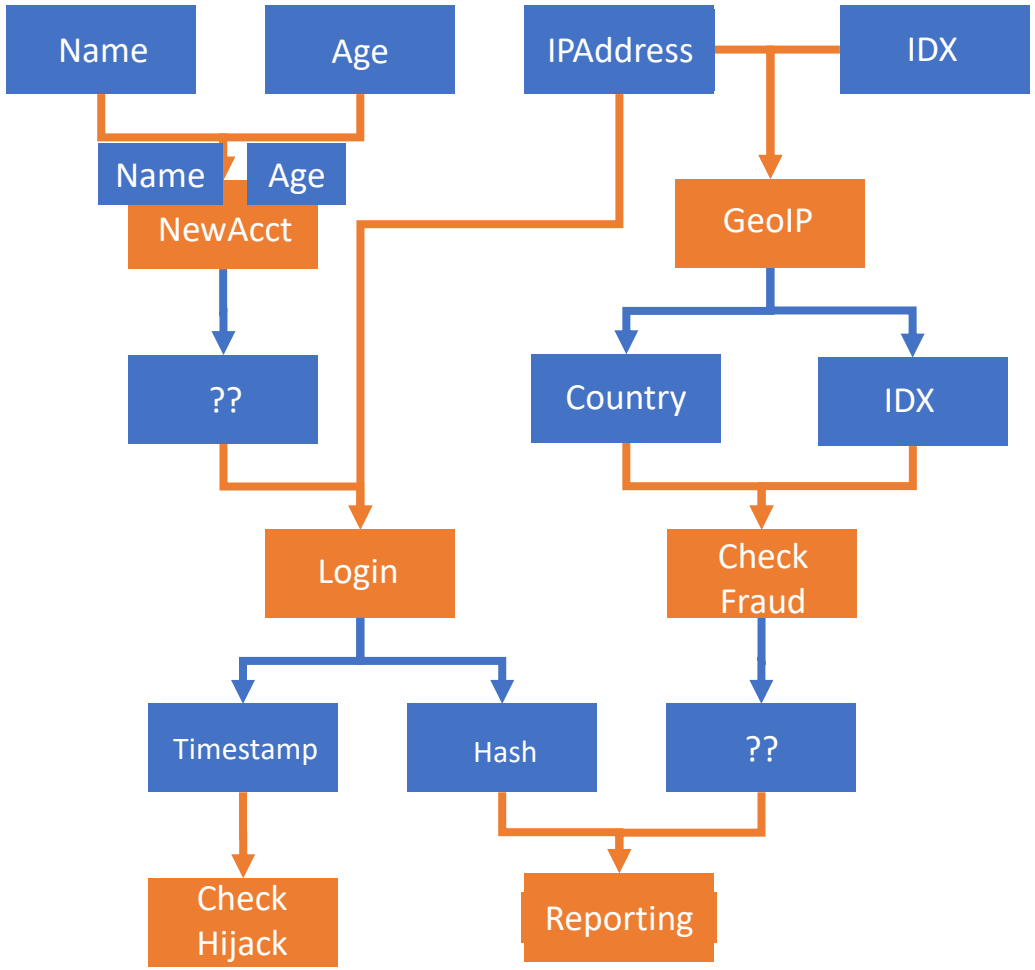
Grok

Purpose Labels

Annotate programs with purpose labels

Initial Data Labels

Heuristics and Annotations



Grok

Purpose Labels

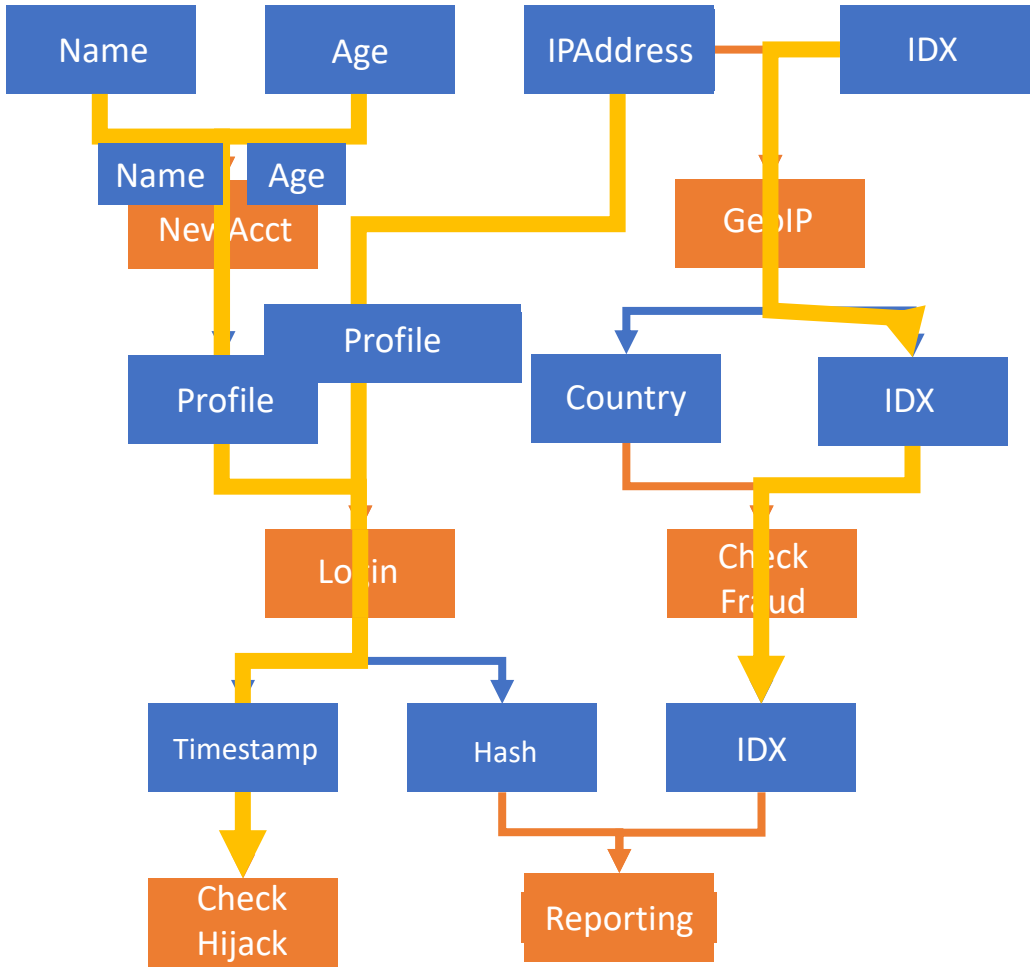
Annotate programs with purpose labels

Initial Data Labels

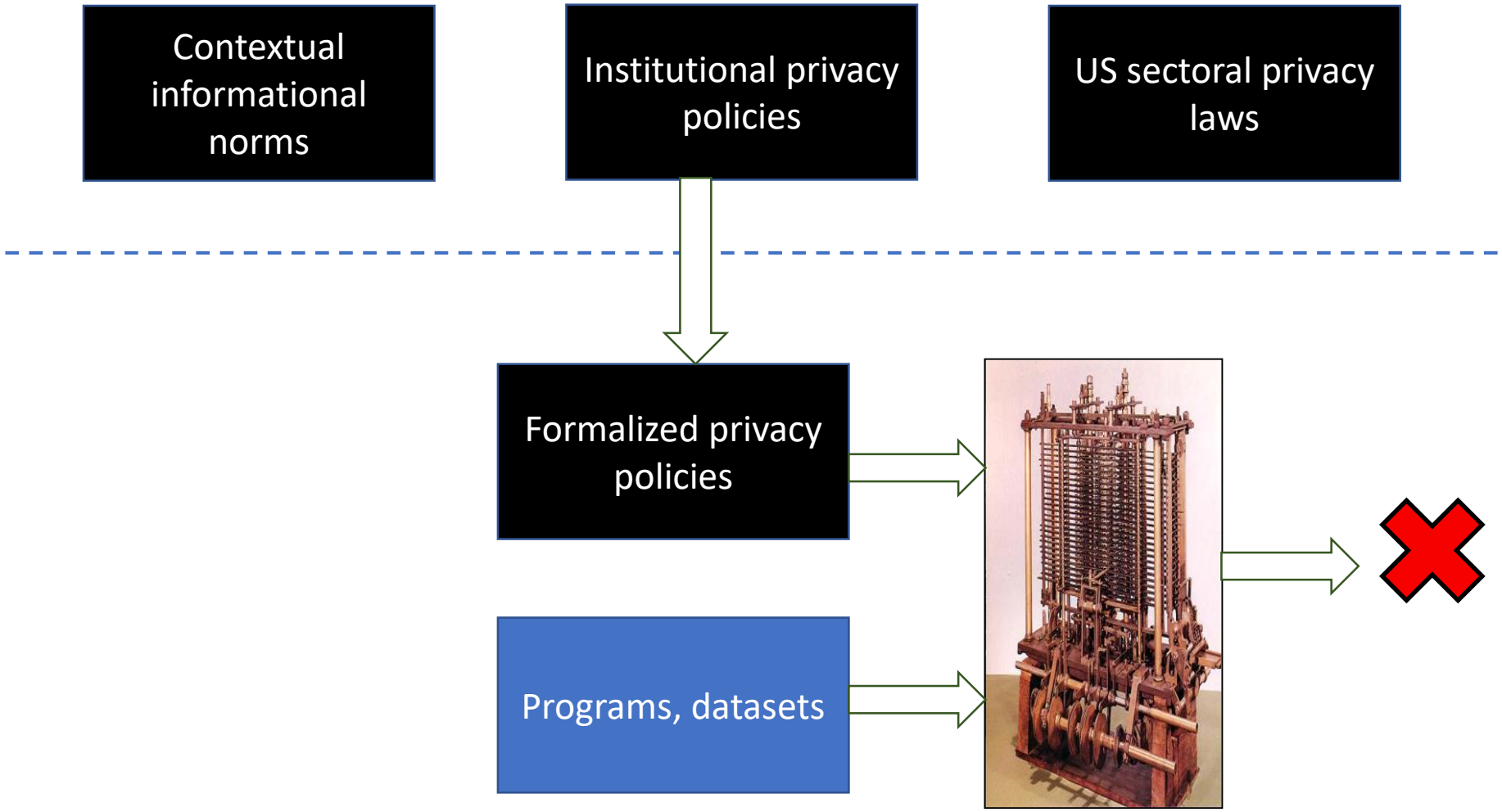
Heuristics and Annotations

Flow Labels

Source labels propagated via data flow graph



Enforcing privacy



Bootstrapping Privacy Compliance in Big Data Systems

Shayak Sen*, Saikat Guha†, Anupam Datta*, Sriram K. Rajamani†, Janice Tsai‡ and Jeannette M. Wing‡
*Carnegie Mellon University, Pittsburgh, USA
{shayaks,danupam}@cmu.edu
†Microsoft Research, Bangalore, India
{saikat,sriram}@microsoft.com
‡Microsoft Research, Redmond, USA
{jatsai,wing}@microsoft.com

- Usable policy language, Legalease, inspired by work on specifying HIPAA, GLBA
- Data inventory, Grok, annotates datatypes in and purposes of programs (non-trivial, likely incomplete)
- Automatic static compliance checking of Bing advertising pipeline
- Deployed on Microsoft production systems for Bing
- Policies in use quite far from CI flow norms (GDPR provides opportunities to change that)

Questions relevant to CI

- What is the “type” (or topic) of a piece of data?
- Is it useful to have incomplete enforcement?
- Should we remove all dependence on semantics of data types?
  - Origin privacy [Benthall, Datta, Tschantz PLSC 2017]
  - Differential privacy [Dwork, McSherry, Nissim, Smith TCC 2006]

Use Privacy in Data-Driven Systems

Theory and Experiments with Machine Learnt Programs

Anupam Datta
Carnegie Mellon University

Matt Fredrikson
Carnegie Mellon University

Gihyuk Ko
Carnegie Mellon University

Piotr Mardziel
Carnegie Mellon University

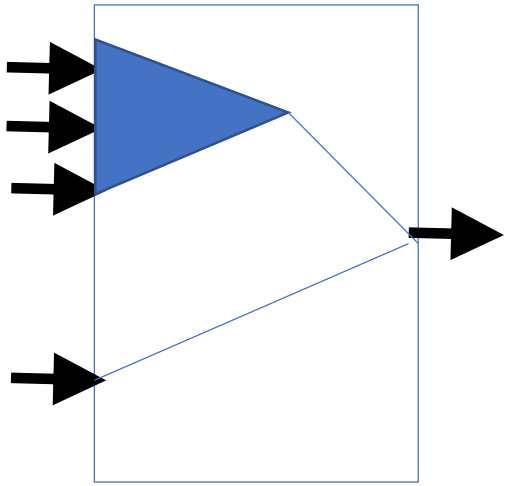
Shayak Sen
Carnegie Mellon University

2017 ACM Conference on Computer and Communications Security

Use Privacy for machine learning models

Protected information type:
Pregnancy status

- Scent-free lotion
- Pre-natal vitamins …



Coupons for diapers?

Proxy use

1. **Strong predictor (associated)**
2. **Causally affects output (high QII)**

Target pregnancy case (2012), Google sleep apnea case (2013-14)

Use Privacy in Data-Driven Systems

Theory and Experiments with Machine Learnt Programs

Anupam Datta
Carnegie Mellon University

Matt Fredrikson
Carnegie Mellon University

Gihyuk Ko
Carnegie Mellon University

Piotr Mardziel
Carnegie Mellon University

Shayak Sen
Carnegie Mellon University

- From epistemic (knowledge) to use restrictions in data-driven systems (beyond CI)
- Indirect use of protected information types outside of expected context

2017 ACM Conference on Computer and Communications Security

Summary

- Contextual integrity is an immensely important piece of the privacy puzzle

Challenges and opportunities

1. What is the “type” (or topic) of a piece of data?
   - Is it useful to have incomplete enforcement?
   - Should we remove all dependence on semantics of data types? (cf. origin privacy, differential privacy)

2. What does it mean to “use” a type of data?
   - Normative theory of use privacy (in addition to epistemic flow-based privacy)
   - Operationalizing use privacy for data-driven systems

3. What does “purpose” mean and how do we enforce purpose restrictions?
   - Initial work in Tschantz, Datta, Wing S&P 2012, ESORICS 2013

4. Deploy in production systems

GDPR, rise of data-driven systems

Princeton, NJ 2018

CENTER FOR INFORMATION TECHNOLOGY POLICY
AT PRINCETON UNIVERSITY

DLI

SYMPOSIUM ON APPLICATIONS OF CONTEXTUAL INTEGRITY

September 13-14, Princeton University.

Co-sponsors:

- Center for Information Technology Policy, Princeton University
- Digital Life Initiative, Cornell Tech.

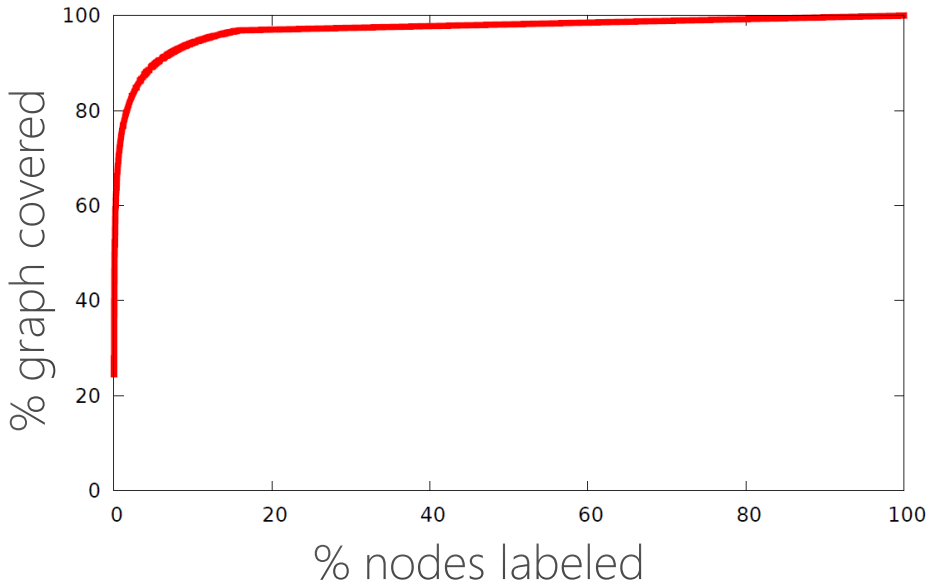
Attendance by invitation-only.

SUSTAINING PRIVACY AND OPEN JUSTICE IN THE TRANSITION TO ONLINE COURT RECORDS: A MULTIDISCIPLINARY INQUIRY[+]

AMANDA CONLEY,[*] ANUPAM DATTA,[**]
HELEN NISSENBAUM[***] & DIVYA SHARMA[****]

- A two-tiered solution? Redacted online version + full version in courthouse
- Finer-grained access rules tied to purpose of accountability of justice system informed by CI?
- Open problem

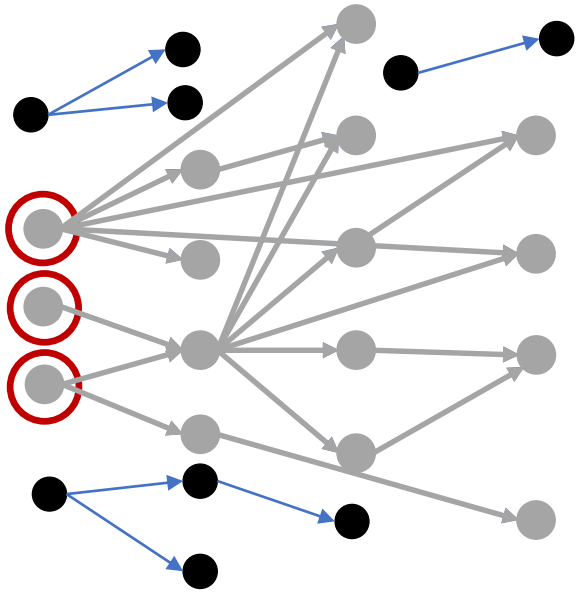
Why Bootstrapping Grok Works



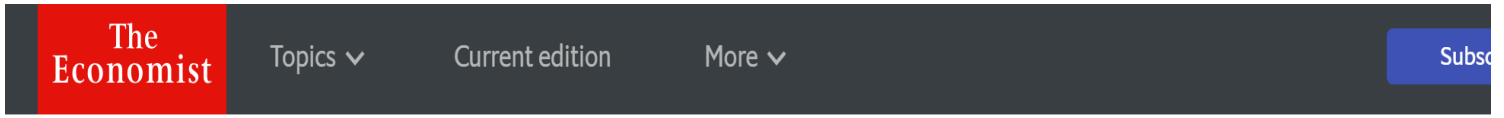
A small number of annotations is enough to get off the ground.

Pick the nodes which will label the most of the graph

~200 annotations label 60% of nodes



Showcased



Personal data

The logic of privacy

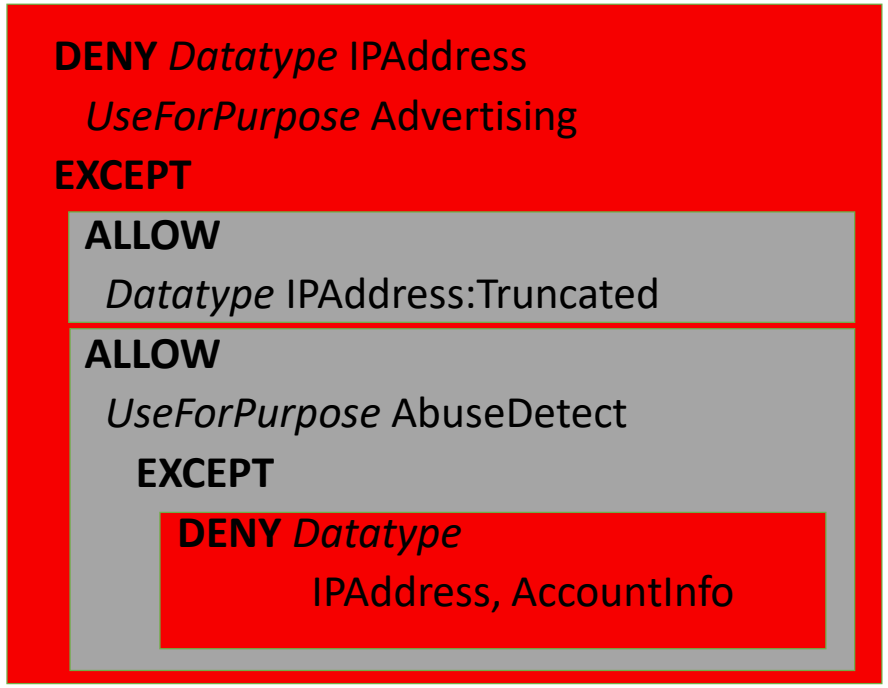
A new way to think about computing and personal information

Print edition | Science and technology ›

Jan 4th 2007

PEOPLE do not have secret trolleys at the supermarket, so how can it be a violation of their privacy if a grocer sells their purchasing habits to a marketing firm? If they walk around in public view, what harm can cameras recording their movements cause? A company is paying them to do a job, so why should it not read their e-mails when they are at work?

Legalease



We will **not** use **full IP Address** for **Advertising**. IP Address may be used for **detecting abuse**. In such cases, it will not be combined with **account information.**