Concurrency Theory meets Security

Based on joint work with:

Kostas Chatzikokolakis, Daniel Gebler, Catuscia Palamidessi, Lili Xu

Focus on:

- Specification and verification of security properties
- Information flow (leakage of confidential information through public observables)
- Quantitative properties

Information leakage in the real world

BBC

News Sport Weather Capital Future Shop

NEWS TECHNOLOGY

Home US & Canada Latin America UK Africa Asia Europe Mid-East Business Health Sci/Environ

SEAMLESS CLOUD FOR THE WORLD
FIND OUT MORE

13 March 2014 Last updated at 21:23 ET

Share

Mark Zuckerberg 'confused and frustrated' by US spying

Mr Zuckerberg said that the internet needed to be made more secure for users

Facebook founder Mark Zuckerberg has said he has called President Barack Obama to "express frustration" over US digital surveillance.

The 29-year-old said in a blog post the US government "should be the champion for the internet, not a threat".

Related Stories

'Spying setting fire to internet'

Trust in the internet

Bits

MARCH 13, 2014, 7:45 AM | Comment

Daily Report: Europe Moves to Reform Rules Protecting Privacy

By THE NEW YORK TIMES

E-MAIL
FACEBOOK
TWITTER
SAVE
MORE

The European Parliament passed a strong new set of data protection measures on Wednesday prompted in part by the disclosure by Edward J. Snowden, a former contractor at the United States National Security Agency, of America's vast electronic spying program, David Jolly reports.

Daily Report

theguardian

News | US | World | Sports | Comment | Culture | Business | Money

News › Society › NHS

NHS England patient data 'uploaded to Google servers', Tory MP says

Health select committee member Sarah Wollaston queries how data was secured by PA Consulting and uploaded to servers outside UK

Police will have 'backdoor' access to health records

Target says it declined to act on early alert of cyber breach

BY JIM FINKLE AND SUSAN HEAVEY

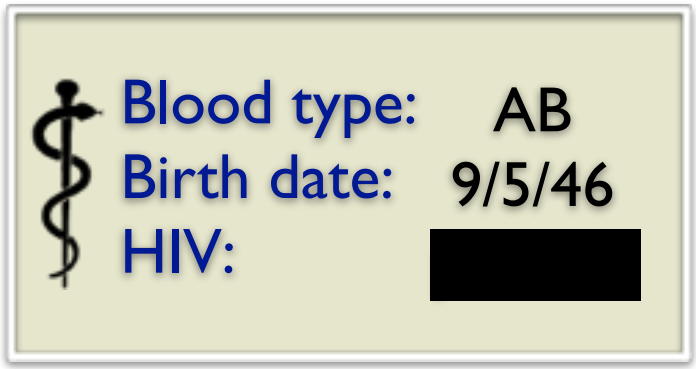
BOSTON/WASHINGTON Thu Mar 13, 2014 6:39pm EDT

5 COMMENTS | Tweet 45 | Share 21 | Share this | 8+1 12 | Email | Print

Merchandise baskets are lined up outside a Target department store in Palm Coast, Florida, December 9, 2013.
CREDIT: REUTERS/LARRY DOWNING

Protection of sensitive information

- Protecting the confidentiality of sensitive information is a fundamental issue in computer security

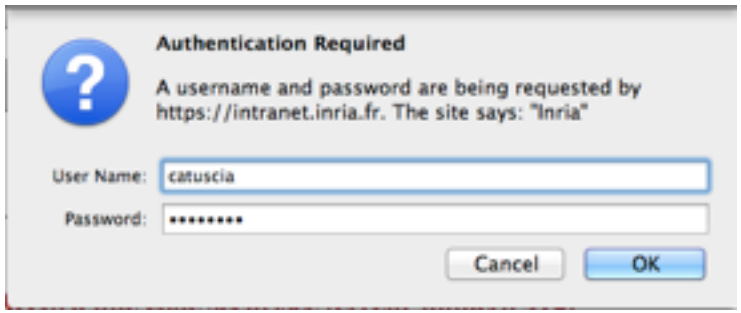


- Access control and encryption are not sufficient! Systems could leak secret information through the correlation with public information.

- The notion of "publicly observable" is subtle and crucial.
  - It may be combined from different sources
  - It may depend on the power of the adversary

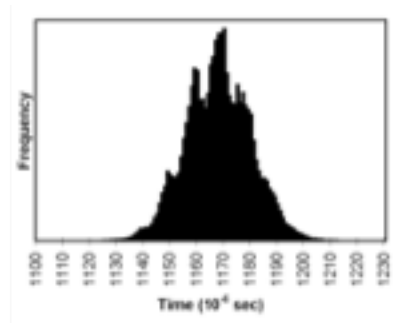
Leakage through correlated observables

Password checking



Election tabulation

Timings of decryptions



Why quantitative properties

1. It is usually impossible to prevent leakage completely. Hence we have to reason about the amount of leakage (probabilistic knowledge)
2. Often methods to protect information use randomization to obfuscate the link between secrets and observables

Randomized methods

An example: Differential Privacy

- Differential privacy [Dwork et al.,2006] is a notion of privacy originated from the area of **Statistical Databases**

- **The problem:** we want to use databases to get statistical information (aka aggregated information), but without violating the privacy of the people in the database

The problem

- The statistical queries should not reveal private information.

- Example: in a database meant to study a certain disease, we may want to ask queries that reveal the correlation between the disease and the age, but we should not be able to derive from this info whether a certain person has the disease.

name	age	disease
Alice	30	no
Bob	30	no
Don	40	yes
Ellie	50	no
Frank	50	yes

Query:
What is the youngest age of a person with the disease?

Answer:
40

Problem:
The adversary may know that Don is the only person in the database with age 40

The problem

- The statistical queries should not unveil private information.

- Example: in a database meant to study a certain disease, we may want to ask queries that reveal the correlation between the disease and the age, but we should not be able to derive from this info whether a certain person has the disease.

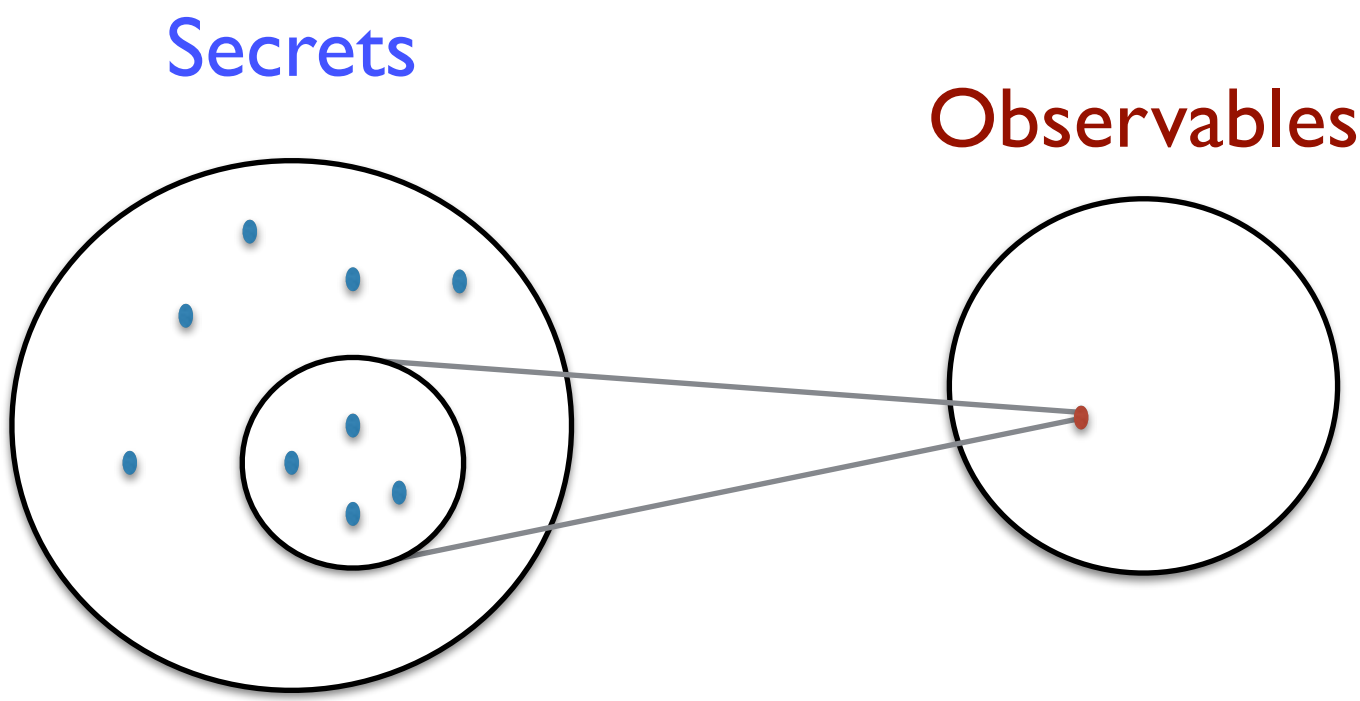
name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

k-anonymity: the answer always partition the space in groups of at least k elements

Alice	Bob
Carl	Don
Ellie	Frank

Many-to-one

- This is a general principle of (deterministic) approaches to protection of confidential information: Ensure that there are **many** secrets that correspond to **one** observable



The problem

Unfortunately, the many-to-one approach is very fragile under **composition**:

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

Alice	Bob
Carl	Don
Ellie	Frank

The problem of composition

Consider the query:

What is the minimal weight of a person with the disease?

Answer: 100

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

The problem of composition

Combine with the two queries:

minimal weight and the minimal age of a person with the disease

Answers: 40, 100

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

Solution

Introduce some probabilistic noise on the answer, so that the answers of minimal age and minimal weight can be given also by other people with different age and weight

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

Noisy answers

minimal age:

40 with probability 1/2

30 with probability 1/4

50 with probability 1/4

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

Alice	Bob
Carl	Don
Ellie	Frank

Noisy answers

minimal weight:
100 with prob. 4/7
90 with prob. 2/7
60 with prob. 1/7

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

Noisy answers

Combination of the answers

The adversary cannot tell for sure whether a certain person has the disease

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

Differential Privacy

- **Differential Privacy** [Dwork 2006]: a randomized mechanism $\mathcal{K}$ provides ε-differential privacy if for all adjacent databases x, x', and for all $z \in \mathcal{Z}$, we have

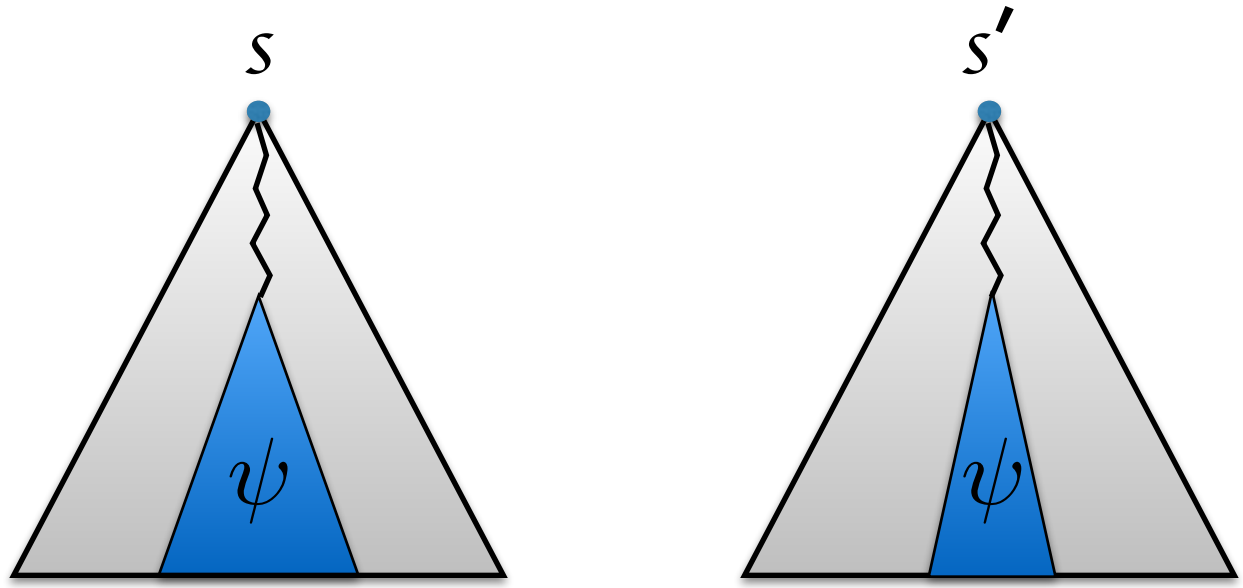
$$\frac{p(K = z|X = x)}{p(K = z|X = x')} \leq e^{\epsilon}$$

- The idea is that the likelihoods of x and x' are not too far apart, for every S
- Equivalent to: learning z changes the probability of x at most by a factor e^{ϵ}
- Differential privacy is robust with respect to composition of queries
- The definition of differential privacy is independent from the prior (but this does not mean that the prior doesn't help in breaching privacy!)
- For certain queries there are mechanisms that are universally optimal, i.e. they provide the best trade-off between privacy and utility, for any prior and any (anti-monotonic) notion of utility

Verification

- We are interested in specifying and verifying quantitative information flow properties in concurrent systems
- Representation:
  - Concurrent systems as probabilistic processes
  - Observables as (observable) traces
  - Secrets as states
- In general, the properties we want to specify and verify are expressed in terms of probabilities of sets of traces
- Hence, a natural approach is to use (bisimulation) metrics. (See the talk by Kim Larsen this morning)
- However, in general information flow properties **are not linear !**

Example: Differential privacy



$$\log \frac{p(s \models \psi)}{p(s' \models \psi)} \leq \epsilon$$

Problems with standard metrics

- Typical properties in quantitative information flow are not linear
  - differential privacy is only an example; the modern approaches to qif are based on information theory and are far from linear
- Hence, the typical metric approaches considered in CT so far are not suitable to specify / verify these properties
  - For example, there can be processes that have finite Kantorovich distance and are not $\in$-differentially private for any $\in$
- However, most qif properties can be expressed in terms of pseudo-distances between the secrets.
  - For example, $\lambda s, s'. \sup_{\psi} \log \frac{p(s \models \psi)}{p(s' \models \psi)}$ (dp) is a pseudo-distance
- Research direction: a parametric pseudometric framework for the specification/verification of distance-based qif properties

Desiderata

- The framework m_d should be parametric wrt the distance d used to express qif properties
- $m_d(s,s') \geq d(s,s')$
- $m_d(s,s')$ should be a bisimulation metric (useful for verification)
- the typical operators should be non-expansive wrt m_d (useful for compositional verification and for reasoning about it)

Progress so far and open problems

- We have a generalized version of the Kantorovich metric that satisfies the desiderata.
- We don't have a general dual form that would allow us to compute the metric easily (but we have it in the case of differential privacy)
- We don't have an elegant solution to integrate the notion of restricted scheduler with a bisimulation metric (cfr. the talk of Pedro d'Argenio this morning)

Thank you !