Generalized Bisimulation Metrics

Catuscia Palamidessi

Based on joint work with:

Kostas Chatzikokolakis, Daniel Gebler, Lili Xu

Plan of the talk

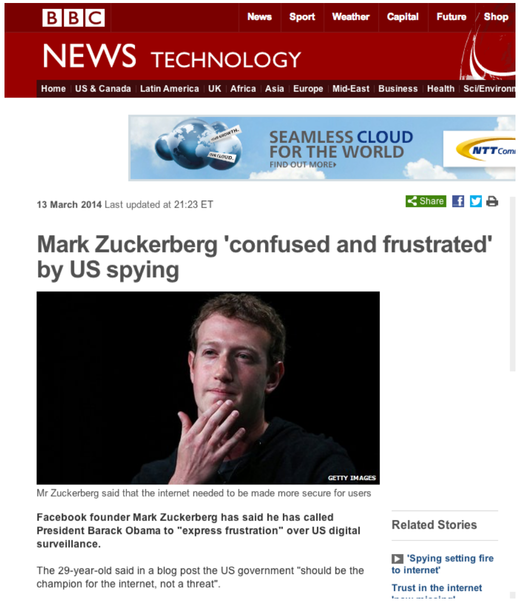
- Motivations
- Desiderata in a notion of pseudo-metric
- Kantorovich metric
- Generalized Kantorovich metric

Motivation

- Formalizing the notion of information leakage in concurrent systems

- Methods for measuring information leakage in a concurrent system and verifying that it is protected against privacy breaches

Information leakage and privacy breaches


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
NTT Com
Share
GETTY IMAGES

13 March 2014 Last updated at 21:23 ET

Mark Zuckerberg 'confused and frustrated' by US spying

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

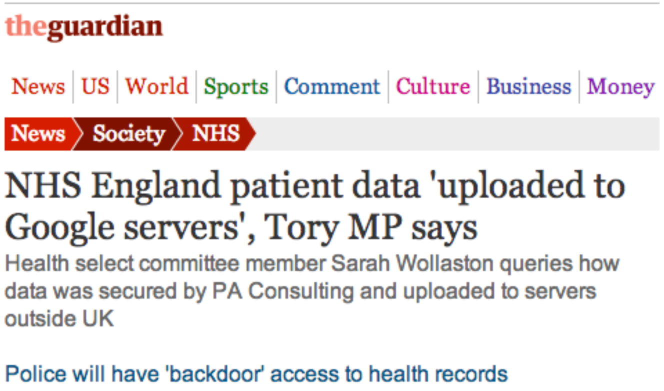
Facebook founder Mark Zuckerberg has said he has called President Barack Obama to "express frustration" over US digital surveillance.

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'

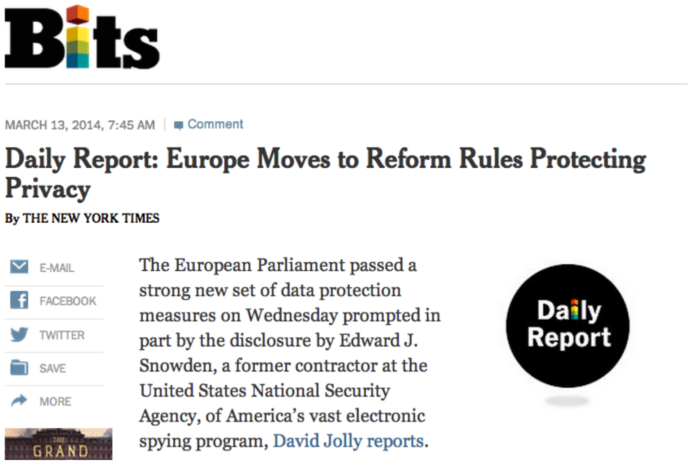
Trust in the internet


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

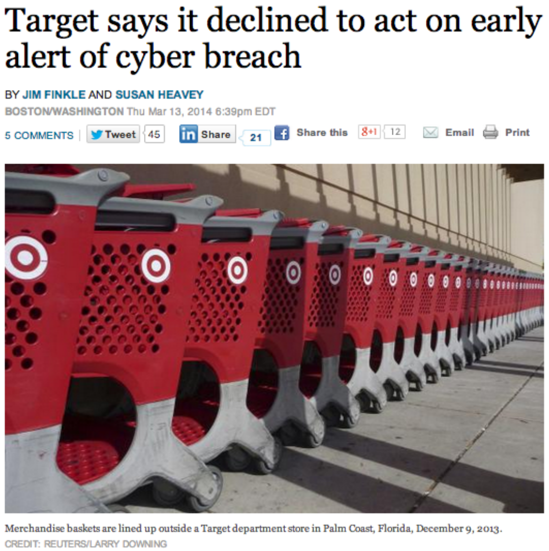


MARCH 13, 2014, 7:45 AM | Comment

Daily Report: Europe Moves to Reform Rules Protecting Privacy

By THE NEW YORK TIMES

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.



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

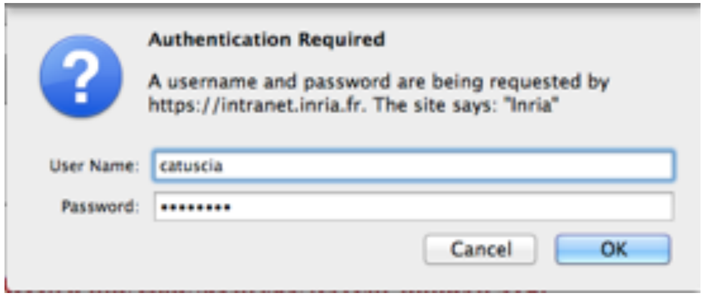
5 COMMENTS

Merchandise baskets are lined up outside a Target department store in Palm Coast, Florida, December 9, 2013.

CREDIT: REUTERS/LARRY DOWNING

Leakage via correlated observables

- Protecting sensitive information is one of the fundamental issues in computer security.



Authentication Required

A username and password are being requested by https://intranet.inria.fr. The site says: "Inria"

User Name: catuscia

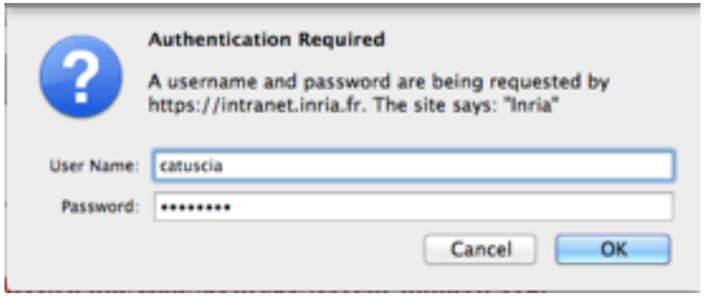
Password: •••••••••

- In several cases Encryption and Access Control can be very effective. However, in this talk we focus in the case in which the leakage of secret information happens through the correlation with public information. This requires a different approach.

- 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

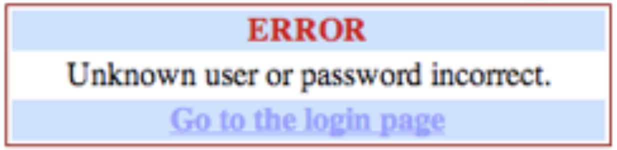


Authentication Required

A username and password are being requested by https://intranet.inria.fr. The site says: "Inria"

User Name: catuscia

Password: •••••••••



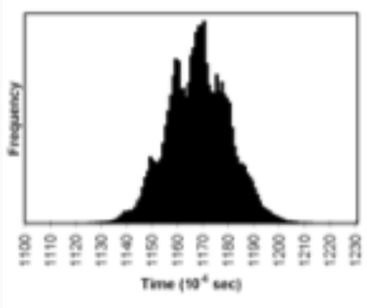
ERROR

Unknown user or password incorrect.

Go to the login page

Election tabulation

Timings of decryptions



Focus on **Quantitative** information leakage

1. It is usually impossible to prevent leakage completely. Hence we need a quantitative notion of leakage. It is usually convenient to reason in terms of **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

- Statistical queries should not reveal private information, but it is not so easy to prevent such privacy breach.

- Example: in a medical database, we may want to ask queries that help to figure the correlation between a disease and the age, but we want to keep private the 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

- Statistical queries should not reveal private information, but it is not so easy to prevent such privacy breach.

- Example: in a medical database, we may want to ask queries that help to figure the correlation between a disease and the age, but we want to keep private the info whether a certain person has the disease.

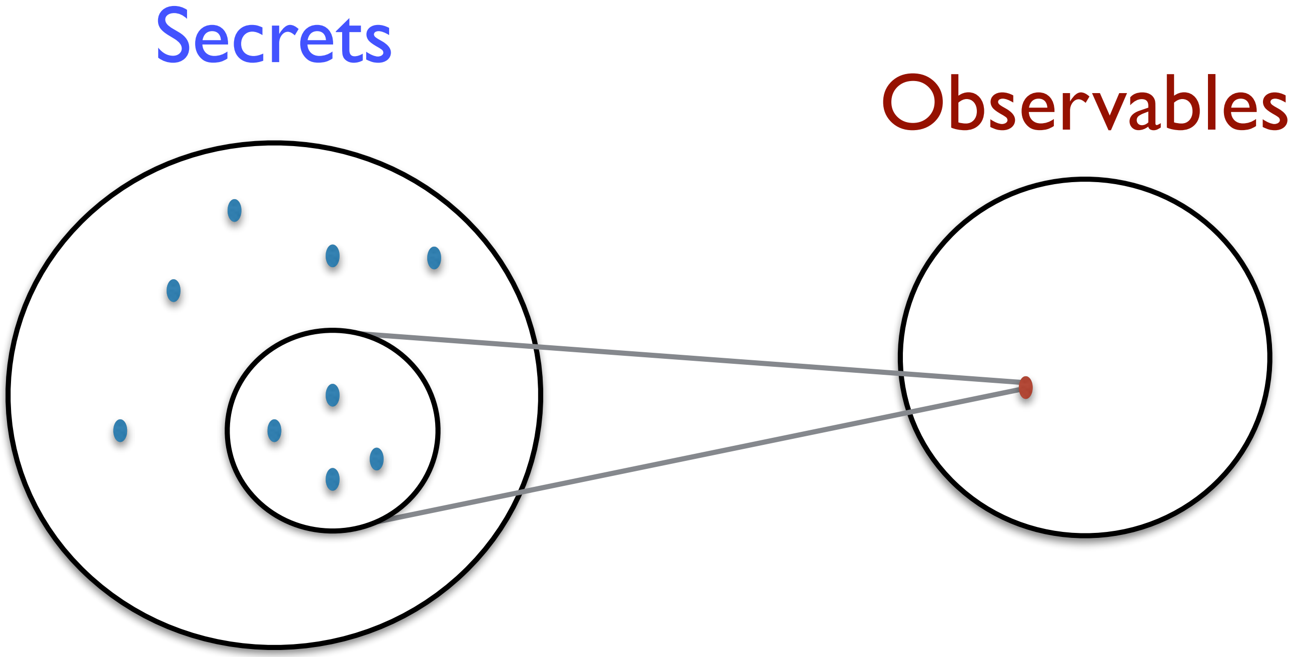
name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

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

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

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

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 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

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

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

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

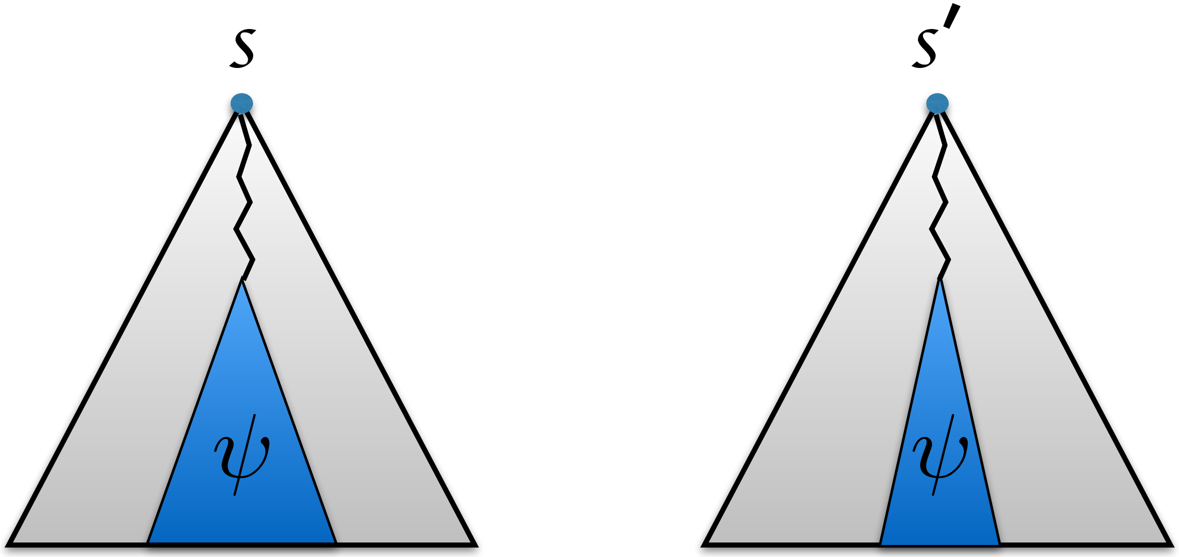
QIF in concurrency

- 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

Example: Differential privacy



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

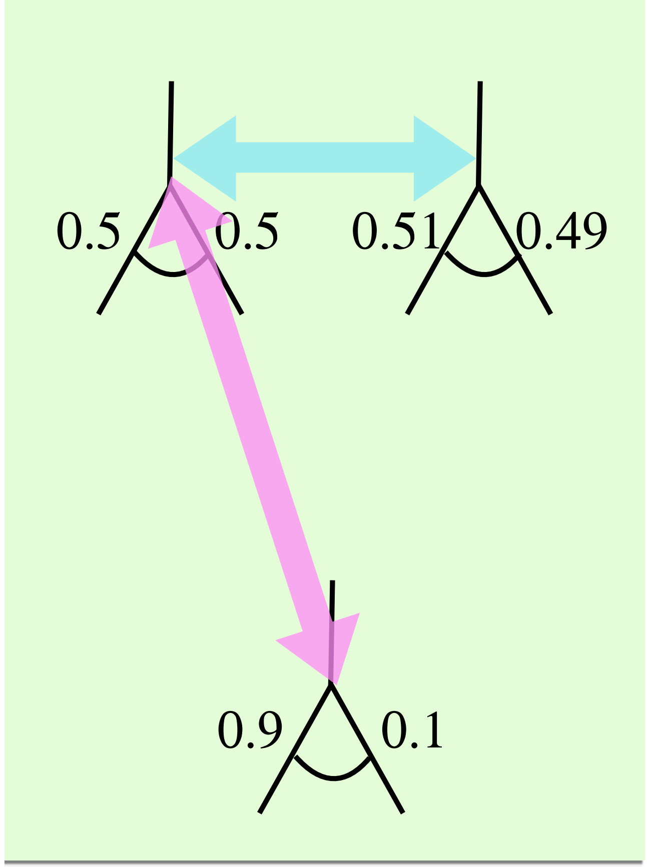
Note that this is a notion of pseudo distance between s and s'

QIF in concurrency

- We need a notion that has good properties and that allows to derive conclusions about traces. In classical process algebra this role is typically played by bisimulation.

From bisimulations to bisimulation metrics

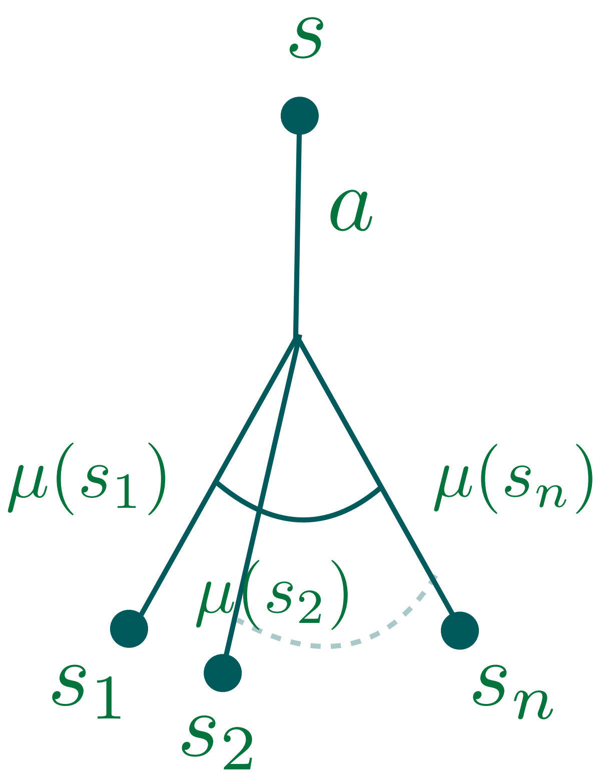
- Bisimulation is a key concept in standard concurrency theory
- However when processes are probabilistic, bisimulation is not robust with respect to small changes of probabilities
- Pseudo distances seems more suitable



Notation

$$s \xrightarrow{a} \mu$$

where s is a state, a is an action,
and μ is a probability distribution



$d(s, s')$:   the distance between s, s'

$d(\mu, \mu')$:   the distance between μ, μ'

Desiderata I

Bisimulation is a well-understood notion, with associated a rich conceptual framework and useful notions and tools, hence we are interested in pseudo metrics that are:

1. conservative extensions of the notion of bisimulation:

$$d(s,s') = 0 \;\text{ iff }\; s \sim s'$$

2. defined via the same kind of coinductive definition, i.e., as greatest fixpoints of the same kind of operator

$$\begin{array}{l} \text{if } \; d(s,s') < \varepsilon \; \text{ then} \\ \quad \text{if } \; s \xrightarrow{a} \mu \; \text{ then } \; \exists \mu' \; \text{ s.t. } \; s' \xrightarrow{a} \mu' \; \text{ and } \; d(\mu,\mu') < \varepsilon \\ \quad \text{if } \; s' \xrightarrow{a} \mu' \; \text{ then } \; \exists \mu \; \text{ s.t. } \; s \xrightarrow{a} \mu \; \text{ and } \; d(\mu,\mu') < \varepsilon \end{array}$$

Desiderata II

3. The typical process algebra operators should be non-expansive wrt the pseudo metric. This is the metric counterpart of the congruence property, and it is useful for compositional reasoning and verification:

$$d(op(s, s_1), op(s, s_2)) \leq d(s_1, s_2)$$

Note: Maybe we could be happy with a weaker property that would only require the expansion to be bound.

4. The pseudo metric should be stronger than the one which defined the QIF property:

$$d'(s, s') \leq d(s, s')$$

where d' is the metric used to define the QIF property

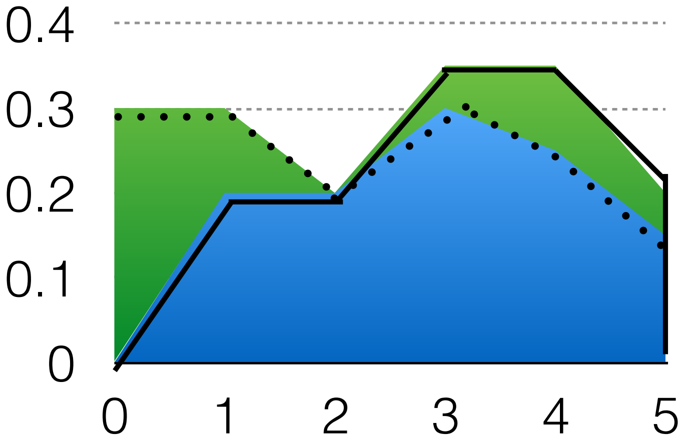
What distance between distributions?

Consider again the formula that defines the pseudo metric coinductively:.

$$
\begin{array}{l}
\text{if } d(s,s') < \varepsilon \text{ then} \\
\quad \text{if } s \xrightarrow{a} \mu \text{ then } \exists \mu' \text{ s.t. } s' \xrightarrow{a} \mu' \text{ and } d(\mu,\mu') < \varepsilon \\
\quad \text{if } s' \xrightarrow{a} \mu' \text{ then } \exists \mu \text{ s.t. } s \xrightarrow{a} \mu \text{ and } d(\mu,\mu') < \varepsilon
\end{array}
$$

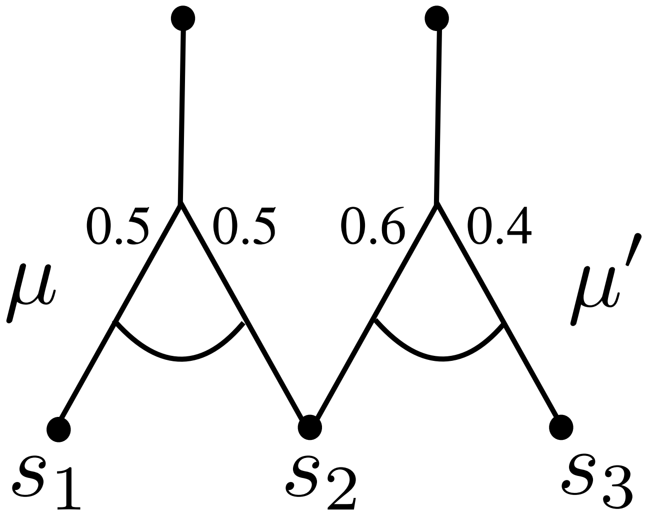
In order to do the coinductive step, we need to lift d from states to distributions on states.

In literature there are several notions of distance between distributions. Typical definitions are those based on the integration of the difference or some norm of the difference



What distance between distributions?

- However, the simple difference between distributions would not make the link between the distances in the coinductive step



- The distance between the two distributions would be the same independently from the distance between s_1 and s_3

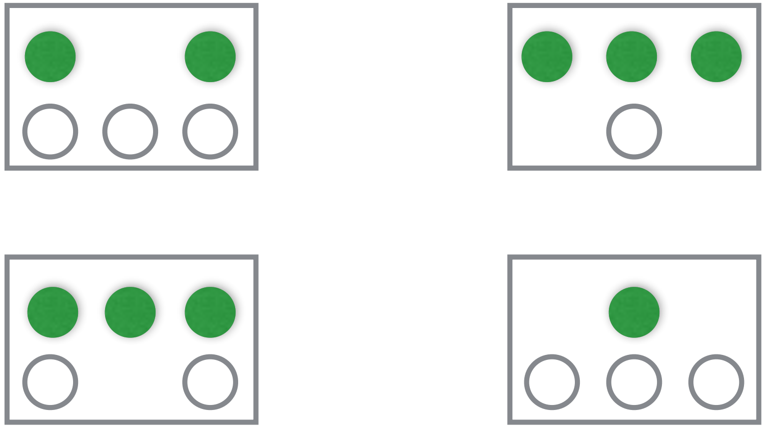
The Kantorovich distance

- The Kantorovich metric allows us to get the proper lifting suitable for the coinductive definition:

$$d(\mu, \mu') = \min_{\alpha} \sum_{s,s'} \alpha(s, s')d(s, s')$$

$$\text{where } \alpha \sum_{s'} \alpha(s, s') = \mu(s) \text{ and } \sum_{s} \alpha(s, s') = \mu'(s')$$

- Transportation problem:



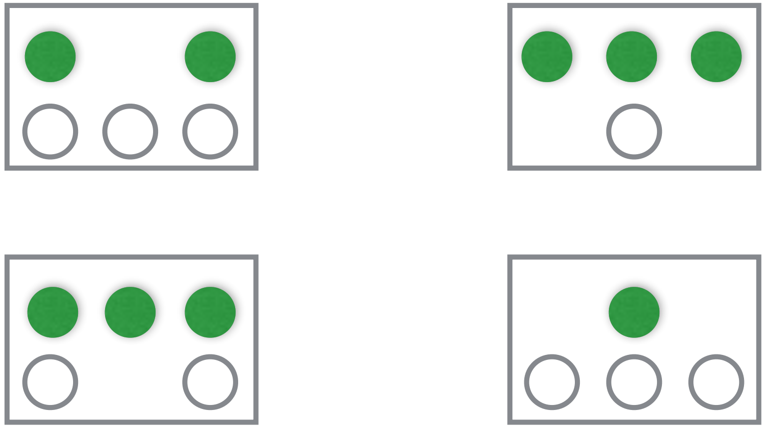
The Kantorovich distance

- The Kantorovich metric allows us to get the proper lifting suitable for the coinductive definition:

$$d(\mu, \mu') = \min_{\alpha} \sum_{s,s'} \alpha(s,s') d(s,s')$$

$$\text{where } \alpha \sum_{s'} \alpha(s,s') = \mu(s) \text{ and } \sum_{s} \alpha(s,s') = \mu'(s')$$

- Transportation problem:



Problems with standard K. metric

- 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 ∈-differentially private for any ∈
- 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

Dual form of the K. metric

$$d(\mu, \mu') = \sup_f \; |\sum_s f(s)\mu(s) - \sum_s f(s)\mu'(s)|$$

Generalization of the K. metric

- In the dual form we substitute the standard difference between reals with the distance that we need for the definition of the QIF property. Let *d'* be this distance. Define:

$$d'(\mu,\mu') = \sup_f \; d'\Big(\sum_s f(s)\mu(s), \sum_s f(s)\mu'(s)\Big)$$

- We have proved that this definition satisfies all the desiderata. In particular, it allows a coinductive construction of a metric that is stronger than the original one of the QIF definition:

- For instance, in the case of differential privacy, we have:

$$d'(\mu,\mu') = \sup_f \log \frac{\sum_s f(s)\mu(s)}{\sum_s f(s)\mu'(s)}$$

Summary and open problems

- We have a generalized version of the Kantorovich metric that satisfies the four desiderata.
- We don't have a general dual form of the "transportation problem" kind that would allow us to compute the metric easily. However we have it in the case of the multiplicative version, corresponding to differential privacy.
- We can handle nondeterminism in the usual way (lifting to the Hausdorff metric), but from the point of view of QIF, unrestricted nondeterminism is problematic. We don't have yet an elegant solution to integrate the notion of restricted scheduler with a bisimulation metric.