A Multiplicative Variant of the Kantorovich distance for Differential Privacy

Lili Xu[1,3,4,5], Konstantinos Chatzikokolakis[2,3],

Catuscia Palamidessi[1,3]

[1] INRIA [2] CNRS [3] Ecole Polytechnique
[4] Grad. Univ. [5] Inst. of Software, Chinese Acad. of Sci.

Originally proposed for privacy protection in the context of statistical databases, differential privacy is now widely adopted in various models of computation. In this paper we investigate techniques for proving differential privacy in the context of concurrent systems containing both probabilistic and non-deterministic behavior. We consider a pseudometric on probabilistic automata which is inspired by the Kantorovich-based bisimulation pseudometric proposed by Desharnais et al. We cannot adopt this notion directly because it does not imply differential privacy. Thus we propose a multiplicative variant of it, which can be characterized in the form of programming problem as well as 1-Lipschitz function, shedding light on the extension to the standard Kantorovich distance. We prove that this multiplicative variant is still an extension of weak bisimulation. Finally we show that the level of differential privacy is continuous on the distance between the starting states in the pseudometric, which makes it suitable for verification.

Keywords: differential privacy, probabilistic automata, bisimulation pseudometrics, verification.

1 Motivation

Differential privacy [7] was originally proposed for privacy protection in the context of statistical databases, but nowadays it is becoming increasingly popular in many other fields, ranging from programming languages [12] to social networks [10] and geolocation [9]. One of the reasons of its success is its independence from side knowledge, which makes it robust to attacks based on combining various sources of information.

In the original definition, a query mechanism $\mathscr{A}$ is ε-differentially private if for any two databases u_1 and u_2 which are adjacent, i.e., differ only for one individual, and any property Z, the probability distributions of $\mathscr{A}(u_1), \mathscr{A}(u_2)$ differ on Z at most by e^{ε}. Namely, $\Pr[\mathscr{A}(u_1) \in Z] \leq e^{\varepsilon} \cdot \Pr[\mathscr{A}(u_2) \in Z]$. This means that the presence (or the data) of an individual cannot be revealed by querying the database. In [4], the principle of differential privacy has been formally extended to measure the degree of protection of secrets in more general settings.

In this paper we deal with the problem of verifying differential privacy properties for concurrent systems, modeled as *probabilistic automata* admitting both nondeterministic and probabilistic behavior. The property of differential privacy requires that the observations generated by two different adjacent secret values be probabilistically similar. In standard concurrent systems the notion of similarity is usually formalized as an equivalence, preferably preserved under composition, i.e., a congruence. Process equivalences have been extensively used to formalize security properties like secrecy [1] and noninterference [8, 13, 14].

In probabilistic systems, we need notions which are robust with respect to small variations in the probabilities, and therefore we usually prefer metric notions over equivalences. In their seminal work,

Submitted to:
QAPL 2014

	Kantorovich metric	The multiplicative variant
Primal	maximize $\sum_i(\mu(s_i)-\mu'(s_i))x_i$ subject to $\forall i.\ 0\le x_i\le 1$ $\forall i,j.\ x_i - x_j \le m(s_i,s_j)$	maximize $\lvert\ln\frac{\sum_i \mu(s_i)x_i}{\sum_i \mu'(s_i)x_i}\rvert$ subject to $\forall i.\ 0\le x_i\le 1$ $\forall i,j.\ x_i \le e^{m(s_i,s_j)}x_j$
Dual	minimize $\sum_{i,j} l_{ij}m(s_i,s_j)+\sum_i x_i+\sum_j y_j$ subject to $\forall i.\ \sum_j l_{ij}+x_i=\mu(s_i)$ $\forall j.\ \sum_i l_{ij}+y_j=\mu'(s_j)$ $\forall i,j.\ l_{ij},x_i,y_j\ge 0$	minimize $\ln z$ subject to $\forall i.\sum_j l_{ij}-r_i=\mu(s_i)$ $\forall j.\ \sum_i l_{ij}e^{m(s_i,s_j)}-r_j\le z\cdot\mu'(s_j)$ $\forall i,j.\ l_{ij},r_i\ge 0$
1-Lipschitz function	$\widehat{m}$ is the smallest metric satisfying: $\forall f: \lvert f(s)-f(s')\rvert\le m(s,s')$ $\Rightarrow \lvert f(\mu)-f(\mu')\rvert\le\widehat{m}(\mu,\mu')$	$\widehat{m}$ is the smallest metric satisfying: $\forall f: \lvert\ln f(s)-\ln f(s')\rvert\le m(s,s')$ $\Rightarrow \lvert\ln f(\mu)-\ln f(\mu')\rvert\le\widehat{m}(\mu,\mu')$

Figure 1: $\widehat{m}(\mu,\mu')$ in the Kantorovich metric and its multiplicative variant.

Desharnais et al. [6] proposed a pseudometric based on the Kantorovich metric, which is particularly appealing because it extends weak bisimilarity (captured by the property of having distance 0) and it is based on a natural way of relating probability masses distributed on a metric space. It also satisfies the property that the composition does not increase the distance, which can be considered the metric generalization of the congruence property.

In order to capture the degree of differential privacy, it is therefore natural to explore also the use of bisimulation metrics, and to consider the metric à la Kantorovic proposed in [6], which represents a cornerstone in this area. However, we cannot use directly the metric of [6] because it does not imply differential privacy: the problem is that the difference in probabilities in this metric is accounted for additively, while differential privacy is a property about their ratio. Thus, we propose a multiplicative variant of it, and obtain a pseudometric that, to the best of our knowledge, is new.

2 A multiplicative variant of the Kantorovich distance

In this section we present a multiplicative variant of the Kantorovich distance in three different characterizations, juxtaposed with the counterparts of the standard Kantorovich distance, displaying clearly the correlation between them.

The Kantorovich metric is a widely used construction for "lifting" a distance from a set of states to distributions over this set. Briefly speaking, it provides a way of measuring the distance between two distributions. Given a set S, we denote by $Disc(S)$ the set of discrete sub-probability measures over S. Let m be a pseudometric on states in S, and $\widehat{m}$ be a pseudometric on distributions in $Disc(S)$. The standard notion of the Kantorovich metric is shown in the left column of Fig. 1 in three forms. Our multiplicative variant is placed in the right column, explained in detail as follows.

Three characterizations

- Primal problem: Initially our pseudometric is obtained by replacing with *the ratio* the statistical difference in the objective function and the constraints in the primal problem of the standard Kantorovich

metric, which turns out to be sufficient to prove the differential privacy property (introduced later).

- Dual problem: Although the primal form can be used to ensure differential privacy, it is not a linear programming problem, thus unpleasant from the point of view of computability. We observe that since ln is a monotonically increasing function, the primal problem is actually a linear-fractional program. It can be converted to an equivalent linear programming problem and therefore a dual program as shown in Fig. 1. This characterization provides a way to compute the distance by using linear programming solution method.
- 1-Lipschitz function characterization: The standard Kantorovich distance has a characterization in terms of 1-Lipschitz function [11]. We consider its discrete setting. Given two metric spaces (S,m) and $(\mathbb{R},d)$, where d denotes the metric on set $\mathbb{R}$, a function $f : S \to \mathbb{R}$ is called 1-Lipschitz continuous if for all s and s' in S,

  $$d(f(s),f(s')) \leq m(s,s').$$

  For simplicity, we overload f on $Disc(s)$ and define $f(\mu) = \sum_{i \in supp(\mu)} \mu(s_i) f(s_i)$.

  Let $d(x,y) = |x-y|$, then the standard Kantorovich metric $\widehat{m}$ is the smallest metric satisfying

  $$d(f(\mu),f(\mu')) \leq \widehat{m}(\mu,\mu')$$

  where f is an arbitrary 1-Lipschitz function with respect to d.

  We find as well a characterization for our multiplicative variant in the form of 1-Lipschitz function. Consider a metric space $(\mathbb{R}^+,d_v)$. Let $d_v(x,y) = |\ln x - \ln y|$ (which can be proved to be a metric), then our multiplicative variant $\widehat{m}$ is the smallest metric satisfying

  $$d_v(f(\mu),f(\mu')) \leq \widehat{m}(\mu,\mu')$$

  where f is an arbitrary 1-Lipschitz function with respect to d_v.

  The primal linear program is actually the problem for obtaining the smallest required pseudometric in the 1-Lipschitz function characterization, meaning that the three forms are equivalent.

Pseudometric on states Based on this multiplicative variant Kantorovich distance, we can define a pseudometric on states of probabilistic automata following the standard line in [6]. We show that our variant satisfies most of the properties of the metric in [6]: in particular, it can be characterized by a fixed-point construction, and it extends weak bisimilarity.

3 Verification of differential privacy

Finally we show that our pseudometric is suitable for verifying differential privacy. Namely, the distance in the pseudometric between two processes determines an upper bound on the ratio of the probabilities of the respective observables.

In concurrent systems, reasoning about the probabilities requires *solving* the nondeterminism first, and to such purpose the usual technique is to consider functions, called *schedulers*, which select the next step based on the history of the computation. However, in our context, as well as in security in general, we need to restrict the power of the schedulers and make them unable to distinguish between secrets in the histories, or otherwise they would plainly reveal them by their choice of the step. See for instance [3, 2, 5] for a discussion on this issue. Thus we consider a restricted class of schedulers, called *admissible*

schedulers, following the definition of [2]. Essentially this definition requires that whenever given two *adjacent* states s, s', namely, differing only for the choice for some secret value, then the choice made by the scheduler on s and s' should be consistent, i.e. the scheduler should not be able to make a different choice on the basis of the secret.

Given a state s, an admissible scheduler ζ, a finite trace $\vec{t}$, we denote by $\Pr_\zeta[s \rhd \vec{t}]$ the probability of producing the trace $\vec{t}$ starting from s under the scheduler ζ. The following theorem shows that our multiplicative version of the Kantorovich metric enjoys a key property:

Theorem 3.1 *For any admissible scheduler ζ, finite trace $\vec{t}$, and adjacent states s, s',*

$$\frac{1}{e^{m(s,s')}} \leq \frac{\Pr_\zeta[s \rhd \vec{t}]}{\Pr_\zeta[s' \rhd \vec{t}]} \leq e^{m(s,s')}$$

From the above result, we derive that to test ε-differential privacy it is sufficient to compute the metric:

Corollary 3.2 *If $m(s,s') \leq \varepsilon$ for any adjancent s, s', then the system is ε-differentially private.*

References

[1] Martín Abadi & Andrew D. Gordon (1999): *A Calculus for Cryptographic Protocols: The Spi Calculus*. *Inf. and Comp.* 148(1), pp. 1–70.

[2] Miguel E. Andrés, Catuscia Palamidessi, Ana Sokolova & Peter Van Rossum (2011): *Information Hiding in Probabilistic Concurrent Systems*. *TCS* 412(28), pp. 3072–3089.

[3] Ran Canetti, Ling Cheung, Dilsun Kaynar, Moses Liskov, Nancy Lynch, Olivier Pereira & Roberto Segala (2006): *Task-Structured Probabilistic I/O Automata*. In: *Proc. of WODES*.

[4] Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe & Catuscia Palamidessi (2013): *Broadening the Scope of Differential Privacy Using Metrics*. In: *Privacy Enhancing Technologies*, pp. 82–102. Available at `http://dx.doi.org/10.1007/978-3-642-39077-7_5`.

[5] Konstantinos Chatzikokolakis & Catuscia Palamidessi (2007): *Making Random Choices Invisible to the Scheduler*. In: *Proc. of CONCUR*, *LNCS* 4703, Springer, pp. 42–58.

[6] Josee Desharnais, Radha Jagadeesan, Vineet Gupta & Prakash Panangaden (2002): *The Metric Analogue of Weak Bisimulation for Probabilistic Processes*. In: *Proc. of LICS*, IEEE, pp. 413–422.

[7] Cynthia Dwork (2006): *Differential Privacy*. In: *Automata, Languages and Programming, 33rd Int. Colloquium, ICALP 2006, Proceedings, Part II*, *LNCS* 4052, Springer, pp. 1–12.

[8] Riccardo Focardi & Roberto Gorrieri (2000): *Classification of Security Properties (Part I: Information Flow)*. In: *FOSAD*, pp. 331–396. Available at `http://dx.doi.org/10.1007/3-540-45608-2_6`.

[9] Ashwin Machanavajjhala, Daniel Kifer, John M. Abowd, Johannes Gehrke & Lars Vilhuber (2008): *Privacy: Theory meets Practice on the Map*. In: *Proc. of ICDE*, IEEE, pp. 277–286.

[10] Arvind Narayanan & Vitaly Shmatikov (2009): *De-anonymizing Social Networks*. In: *Proc. of S&P*, IEEE, pp. 173–187, doi:`10.1109/SP.2009.22`.

[11] S. T. Rachev & Ludger Rüschendorf (1998): *Mass Transportation Problems Volume I: Theory*. Springer.

[12] Jason Reed & Benjamin C. Pierce (2010): *Distance makes the types grow stronger: a calculus for differential privacy*. In: *Proc. of ICFP*, ACM, pp. 157–168.

[13] Peter Y. A. Ryan & Steve A. Schneider (2001): *Process Algebra and Non-Interference*. *Journal of Computer Security* 9(1/2), pp. 75–103.

[14] Geoffrey Smith (2003): *Probabilistic Noninterference through Weak Probabilistic Bisimulation*. In: *CSFW*, pp. 3–13. Available at `http://doi.ieeecomputersociety.org/10.1109/CSFW.2003.1212701`.