Probabilistic and nondeterministic aspects of Anonymity [1]

Catuscia Palamidessi [2]

INRIA and LIX
École Polytechnique, Rue de Saclay, 91128 Palaiseau Cedex, FRANCE

Abstract

Anonymity means that the identity of the user performing a certain action is maintained secret. The protocols for ensuring anonymity often use random mechanisms which can be described probabilistically. The user, on the other hand, may be selected either nondeterministically or probabilistically. We investigate various notions of anonymity, at different levels of strength, for both the cases of probabilistic and nondeterministic users.

Key words: Anonymity, Probabilistic Automata, conditional probability.

1 Introduction

Anonymity is the property of keeping secret the identity of the user performing a certain action. The need for anonymity may raise in a wide range of situations, like postings on electronic forums, voting, delation, donations, and many others.

The protocols for ensuring anonymity often use random mechanisms which can be described probabilistically. This is the case, for example, of the Dining Cryptographers [3], Crowds [7], and Onion Routing [12]. In contrast, we usually don't know anything about the users, so their behavior, and in particular, the choice of the user who performs the action with respect to which we want to ensure anonymity, should better be regarded as nondeterministic. (The same would hold for adversaries, although in this paper we do not consider them.) The whole system constituted by the protocol and the users presents therefore both probabilistic and nondeterministic aspects.

[1] This work has been partially supported by the Project Rossignol of the ACI Sécurité Informatique (Ministère de la recherche et nouvelles technologies).

[2] Email: catuscia@lix.polytechnique.fr

This is a preliminary version. The final version will be published in Electronic Notes in Theoretical Computer Science
URL: www.elsevier.nl/locate/entcs

Various formal definitions and frameworks for analyzing anonymity have been developed in literature. They can be classified into approaches based on process-calculi [9,8], epistemic logic [11,5], and "function views" [6]. From the point of view of the concepts of probability and nondeterminism, however, all these approaches are either *purely nondeterministic* (also known as *possibilistic*) or *purely probabilistic*.

The purely nondeterministic approach in [9,8] is based on the so-called "principle of confusion": a system is anonymous if the set of the possible outcomes is saturated with respect to the intended anonymous users, i.e. if one such user can cause a certain observable trace in one possible computation, then there must be alternative computations in which each other anonymous user can give rise to the same observable trace (modulo the identity of the anonymous users).

The purely probabilistic proposals can be classified under two different points of view: those which focus on the probability of the users, and those which focus on the effect that the observables have on the probability of the users. The distinction is subtle but fundamental. In the fist case, anonymity holds when (an observer knows that) all users have the same probability of having performed the action (cfr. *strong probabilistic anonymity* in [5]). In the second case, it holds when the for any user i and any observable o the conditional probability that i has performed the action, given the observable, is the same as the (a priory) probability that the user has performed the action (cfr. the informal notion used in [3], and the *conditional probabilistic anonymity* in [5]).

The probabilistic approach also brings naturally to differentiate the notion of anonymity with respect to different levels of strength. Reiter and Robin [7] have proposed the following hierarchy:

Beyond suspicion The actual user (i.e. the user that performed the action) is not more likely (to have performed the action) than every other user.

Probable innocence The actual user has probability less than 1/2.

Possible innocence There is a non trivial probability that another used could have performed the action.

These notions were only given informally in [7], and it is unclear to us whether the authors had in mind the first or the second of the "points of view" described above. On one hand, if we interpret the informal definitions literally, they correspond to the first point of view. This is the interpretation given by Halpern and O'Neill in [5]: they characterize probable innocence and possible innocence with the notion of (probabilistic) α-anonymity, and beyond suspicion with their notion of strong probabilistic anonymity. On the other hand, the result of probable innocence proved in [7] for Crowds does not seem to fit with this interpretation, while it could fit with a suitable weakening of the anonymity notion illustrated above under the second perspective (i.e. what Halpern and O'Neill call conditional probabilistic anonymity).

In this work we assume that the users may be nondeterministic, i.e. that nothing may be known about the relative frequency by which each user perform the anonymous action. More precisely, the users can in principle be totally unpredictable and change intention every time, so that their behavior cannot be thought of as probabilistic [3]. The internal mechanisms of the systems, on the contrary, like coin tossing in the dining philosophers, or the random selection of a nearby node in Crowds, are supposed to exhibit a certain regularity and obey a probabilistic distribution. Correspondingly, we explore a notion of probabilistic anonymity that focuses on the internal mechanism of the system, i.e. their non-leakage of probabilistic information, and it is in a sense independent from the users in case they are nondeterministic. The counterpart of our definition in the case the users are probabilistic (with possibly unknown probabilities), can be shown to correspond to a generalized version of Halpern and O'Neill's conditional probabilistic anonymity, where "generalized" here means that the anonymity holds for any probability distribution to the users.

This abstract is based on the ongoing work reported in [1], [4] and [2].

2 Nondeterminism and probability

In our approach we consider systems that can perform both probabilistic and nondeterministic choice. Intuitively, a probabilistic choice represents a set of alternative transitions, each of them associated to a certain probability of being selected. The sum of all probabilities on the alternatives of the choice must be 1, i.e. they form a *probability distribution*. Nondeterministic choice is also a set of alternatives, but we have no information on how likely one alternative is selected.

There have been many models proposed in literature that combine both nondeterministic and probabilistic choice. One of the most general is the formalism of *probabilistic automata* proposed in [10]. We give here a brief and informal description of it.

A probabilistic automaton consists in a set of states, and labeled transitions between them. For each node, the outgoing transitions are partitioned in groups called *steps*. Each step represents a probabilistic choice, while the choice between the steps is nondeterministic.

Figure 1 illustrates some examples of probabilistic automata. We represent a step by putting an arc across the member transitions. For instance, in (a),

[3] In the areas of concurrency theory there has been a long-standing discussion on whether nondetermistism can be thought of as a situation in which the probabilities are unknown and can change very time the experiment is repeated. Nowadays the prevailing opinion, which we share, is that nondetermistism and probability are fundamentally different concepts. Furthermore, in the formalisms used in concurrency (for instance process algebras) the difference is clear, as these two concepts (nondeterminism and unknown changing probability) obey different laws.

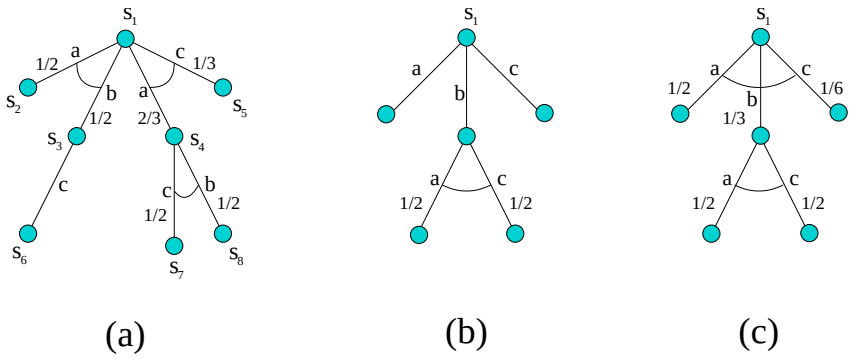

Fig. 1. Examples of probabilistic automata

state s_1 has two steps, the first is a probabilistic choice between two transitions with labels a and b, each with probability $1/2$. When there is only a transition in a step, like the one from state s_3 to state s_6, the probability is of course 1 and we omit it.

In this paper, we use only a simplified kind of automaton, in which from each node we have either a probabilistic choice or a nondeterministic choice (more precisely, either one step or a set of singleton steps), like in (b). In the particular case that the choices are all probabilistic, like in (c), the automaton is called *fully probabilistic*.

Given an automaton M, we denote by $etree(M)$ its unfolding, i.e. the tree of all possible executions of M (in Figure 1 the automata coincide with their unfolding because there is no loop). If M is fully probabilistic, then each execution (maximal branch) of $etree(M)$ has a probability obtained as the product of the probability of the edges along the branch. In the finite case, we can define a probability measure for each set of executions, called *event*, by summing up the probabilities of the elements[4]. Given an event x, we will denote by $p(x)$ the probability of x. For instance, let the event c be the set of all computations in which c occurs. In (c) its probability is $p(c) = 1/3 \times 1/2 + 1/6 = 1/3$.

When nondeterminism is present, the probability can vary, depending on how we *resolve* the nondeterminism. In other words we need to consider a function ς that, each time there is a choice between different steps, selects one of them. By pruning the non-selected steps, we obtain a fully probabilistic execution tree $etree(M, \varsigma)$ on which we can define the probability as before. For historical reasons (i.e. since nondeterminism typically arises from the parallel operator), the function ς is called *scheduler*.

It should then be clear that the probability of an event is relative to the particular scheduler. We will denote by $p_\varsigma(x)$ the probability of the event x under the scheduler ς. For example, consider (a). We have two possible

[4] In the infinite case things are more complicated: we cannot define a probability measure for all sets of execution, and we need to consider as event space the σ-field generated by the *cones* of $etree(M)$. However, in this paper, we consider only the finite case.

schedulers determined by the choice of the step in s_1. Under one scheduler, the probability of c is $1/2$. Under the other, it is $2/3 \times 1/2 + 1/3 = 2/3$. In (b) we have three possible schedulers under which the probability of c is 0, $1/2$ and 1, respectively.

3 Anonymity systems

We model the anonymity protocol as a probabilistic automaton M. The concept of anonymity is relative to the set of anonymous users and to what is visible to the observer. Hence, following [9,8] we classify the actions of M into the three sets A, B and C as follows:

- A is the set of the anonymous actions $A = \{a(i) \mid i \in I\}$ where I is the set of the identities of the anonymous users and a is an injective functions from I to the set of actions, which we call *abstract action*. We also call the pair (I, a) *anonymous action generator*.
- B is the set of the observable actions. We will use b, b', ... to denote the elements of this set.
- C is the set of the remaining actions (which are unobservable).

Note that the actions in A normally are not visible to the observer, or at least, not for the part that depends on the identity i. However, for the purpose of defining and verifying anonymity we model the elements of A as visible outcomes of the system.

Definition 3.1 An anonymity system is a tuple (M, I, a, B, Z, p), where M is a probabilistic automaton, (I, a) is an anonymous action generator, B is a set of observable actions, Z is the set of all possible schedulers for M, and for every $\varsigma \in Z$, p_ς is the probability measure on the event space generated by $\mathit{etree}(M, \varsigma)$.

For simplicity, we assume the users to be the only possible source of nondeterminism in the system. If they are probabilistic, then the system is fully probabilistic, hence Z is a singleton and we omit it.

We introduce the following notation to represent the events of interest:

- $a(i)$: all the executions in $\mathit{etree}(M, \varsigma)$ containing the action $a(i)$;
- a : all the executions in $\mathit{etree}(M, \varsigma)$ containing an action $a(i)$ for an arbitrary i;
- o : all the executions in $\mathit{etree}(M, \varsigma)$ containing as their maximal sequence of observable actions the sequence o (where o is of the form $b_1 b_2 \ldots b_n$ for some $b_1, b_2, \ldots, b_n \in B$). We denote by O (*observables*) the set of all such o's.

We use the symbols $\cup$, $\cap$ and $\neg$ to represent the union, the intersection, and the complement of events, respectively.

We wish to keep the notion of observables as general as possible, but we still need to make some assumptions on them. First, we want the observables to be disjoint events. Second, they must cover all possible outcomes. Third, an observable o must indicate unambiguously whether a has taken place or not, i.e. it either implies a, or it implies $\neg a$. In set-theoretic terms it means that either o is a subset of a or of the complement of a. Formally:

Assumption 1 (on the observables)

(i) $\forall \varsigma \in Z.\ \forall o_1, o_2 \in O.\ \ o_1 \neq o_2 \ \Rightarrow\ p_\varsigma(o_1 \cup o_2) = p_\varsigma(o_1) + p_\varsigma(o_2)$

(ii) $\forall \varsigma \in Z.\ \ p_\varsigma(O) = 1$

(iii) $\forall \varsigma \in Z.\ \forall o \in O.\ \ (p_\varsigma(o \cap a) = p_\varsigma(o)) \quad \vee \quad p_\varsigma(o \cap \neg a) = p_\varsigma(o)$

Analogously, we need to make some assumption on the anonymous actions. We consider first the conditions tailored for the nondeterministic users: each scheduler determines completely whether an action of the form $a(i)$ takes place or not, and in the positive case, there is only one such i. Formally:

Assumption 2 (on the anonymous actions, for nondeterministic users)

$$\forall \varsigma \in Z.\ \ p_\varsigma(a) = 0 \vee (\exists i \in I.\ (p_\varsigma(a(i)) = 1\ \wedge\ \forall j \in I.\ j \neq i \Rightarrow p_\varsigma(a(j)) = 0))$$

We now consider the case in which the users are fully probabilistic. The assumption on the anonymous actions in this case is much weaker: we only require that there be at most one user that performs a, i.e. $a(i)$ and $a(j)$ must be disjoint for $i \neq j$. Formally:

Assumption 3 (on the anonymous actions, for probabilistic users)

$$\forall i, j \in I.\ \ i \neq j\ \Rightarrow\ p(a(i) \cup a(j)) = p(a(i)) + p(a(j))$$

4 Strong probabilistic anonymity

In this section we recall the notion of strong anonymity proposed in [1].

Let us first assume that the users are nondeterministic. Intuitively, a system is strongly anonymous if, given two schedulers ς and ϑ that both choose a (say $a(i)$ and $a(j)$, respectively), it is not possible to detect from the probabilistic measure of the observables whether the scheduler has been ς or ϑ (i.e. whether the selected user was i or j).

Definition 4.1 A system (M, I, a, B, Z, p) with nondeterministic users is anonymous if

$$\forall \varsigma, \vartheta \in Z.\ \forall o \in O.\ p_\varsigma(a) = p_\vartheta(a) = 1\ \Rightarrow\ p_\varsigma(o) = p_\vartheta(o)$$

The probabilistic counterpart of Definition 4.1 can be formalized using the concept of *conditional probability*. Recall that, given two events x and y with $p(y) > 0$, the conditional probability of x given y, denoted by $p(x \mid y)$, is equal to $p(x \cap y)/p(y)$.

Definition 4.2 A system (M, I, a, B, p) with probabilistic users is anonymous if

$$\forall i, j \in I.\ \forall o \in O.\ \ (p(a(i)) > 0 \wedge p(a(j)) > 0) \Rightarrow p(o \,|\, a(i)) = p(o \,|\, a(j))$$

The notions of anonymity illustrated so far focus on the probability of the observables. More precisely, it requires the probability of the observables to be independent from the selected user. In [1] it was shown that Definition 4.2 is equivalent to the notion adopted implicitly in [3], and called *conditional anonymity* in [5]. As illustrated in the introduction, the idea of this notion is that a system is anonymous if the observations do not change the probability of the $a(i)$'s. In other words, we may know the probability of $a(i)$ by some means external to the system, but the system should not increase our knowledge about it.

Proposition 4.3 *A system (M, I, a, B, p) with probabilistic users is anonymous if*

$$\forall i \in I.\ \forall o \in O.\ \ p(o \cap a) > 0 \ \Rightarrow\ p(a(i) \,|\, o) = p(a(i) \,|\, a)$$

The notions of strong anonymity proposed in this section have been verified in [1] on the example of the Dining Cryptographers with perfectly fair coins.

4.1 Independence from the probability distribution of the users

One important property of Definition 4.2 is that it is independent from the probability distribution of the users. Intuitively, this is due to the fact that the condition of anonymity implies that $p(o \,|\, a(i)) = p(o)/p(a)$, hence it is independent from $p(a(i))$.

Theorem 4.4 *If (M, I, a, B, p) is anonymous (according to Definition 4.2) then for any p' which differs from p only on the $a(i)$'s, (M, I, a, B, p') is anonymous.*

Also the characterization of anonymity given in Proposition 4.3 is (obviously) independent from the probability distribution of the users. It should be remarked, however, that their correspondence with Definition 4.2, and the property of independence from the probability of the users, only holds under the hypothesis that there is at most one agent performing a. (Assumption 3.)

5 Weak probabilistic anonymity

The notion of anonymity proposed in previous section is very strong as it imply the system to leak no information. In reality, some amount of probabilistic information may be revealed by the protocol. Typical causes may be either the presence of attackers which interfere with the normal execution the protocol, or some unavoidable imperfection of the internal mechanisms, or may even

be inherent to the way the protocol is designed. In any case, the information leaked by the system can be used by an observer to infer the likeliness that the action has been performed by a certain user.

In [4] we have proposed the following weak variants of Definition 4.1, Definition 4.2, and of the notion expressed in Proposition 4.3.

Let us consider first the case of nondeterministic users:

Definition 5.1 Given $\alpha \in [0,1]$, a system (M, I, a, B, Z, p) with nondeterministic users is α-anonymous if

$$max\{\, p_\varsigma(o) - p_\vartheta(o) \;\mid \varsigma, \vartheta \in Z,\ o \in O,\ p_\varsigma(o \cap a) = p_\varsigma(o),\ p_\vartheta(o \cap a) = p_\vartheta(o)\} = \alpha$$

Intuitively, $p_\varsigma(o) - p_\vartheta(o) = \alpha$ means that, whenever we observe o, we suspect that user i is more likely than user j to have performed the action by an additive factor α (where i and j represent the users selected by ς and ϑ, respectively).

Let us now consider the case of probabilistic users. The weak version of Definition 4.2, is the following:

Definition 5.2 Given $\alpha \in [0,1]$, a system (M, I, a, B, p) with probabilistic users is α-anonymous if

$$max\{\, p(o \mid a(i)) - p(o \mid a(j)) \;\mid i, j \in I,\ o \in O,\ p(a(i)) > 0,\ p(a(j)) > 0\,\} = \alpha$$

An alternative notion of weak anonymity for probabilistic users can be obtained by weakening the formula in Proposition 4.3.

Definition 5.3 Given $\alpha \in [0,1]$, a system (M, I, a, B, p) with probabilistic users is α-anonymous if

$$max\{\, p(a(i) \mid o) - p(a(i) \mid a) \;\mid i \in I,\ o \in O,\ p(o \cap a) > 0\} = \alpha$$

Intuitively, $p(a(i) \mid o) - p(a(i) \mid a) = \alpha$ means that, after observing o, the probability we attribute to i as the performer of the action, has increased by an additive factor α.

Differently from their strong version, Definitions 5.2 and Definitions 5.3 are not equivalent.

The notions of weak anonymity proposed in this section have been tested in [4] on the example of the Dining Cryptographers with biased coins.

In [2] it has been proved that a slight variant of Definitions 5.2 and 5.3 correspond, with $\alpha = 1/2$, to the property proved in [7] for the system Crowds (and called, still in [7], probable innocence).

References

[1] Mohit Bhargava and Catuscia Palamidessi. Probabilistic anonymity. Technical report, INRIA Futurs and LIX, 2005. Submitted for publication.

http://www.lix.polytechnique.fr/~catuscia/papers/Anonymity/report.ps.

[2] Kostantinos Chatzikokolakis and Catuscia Palamidessi. Probable innocence revisited. Technical report, INRIA Futurs and LIX, 2005. http://www.lix.polytechnique.fr/~catuscia/papers/Anonymity/reportPI.pdf.

[3] David Chaum. The dining cryptographers problem: Unconditional sender and recipient untraceability. *Journal of Cryptology*, 1:65–75, 1988.

[4] Yuxin Deng, Catuscia Palamidessi, and Jun Pang. Weak probabilistic anonymity. Technical report, INRIA Futurs and LIX, 2005. Submitted for publication. http://www.lix.polytechnique.fr/~catuscia/papers/Anonymity/reportWA.pdf.

[5] Joseph Y. Halpern and Kevin R. O'Neill. Anonymity and information hiding in multiagent systems. In *Proc. of the 16th IEEE Computer Security Foundations Workshop*, pages 75–88, 2003.

[6] Dominic Hughes and Vitaly Shmatikov. Information hiding, anonymity and privacy: a modular approach. *Journal of Computer Security*, 12(1):3–36, 2004.

[7] Michael K. Reiter and Aviel D. Rubin. Crowds: anonymity for Web transactions. *ACM Transactions on Information and System Security*, 1(1):66–92, 1998.

[8] Peter Y. Ryan and Steve Schneider. *Modelling and Analysis of Security Protocols*. Addison-Wesley, 2001.

[9] Steve Schneider and Abraham Sidiropoulos. CSP and anonymity. In *Proc. of the European Symposium on Research in Computer Security (ESORICS)*, volume 1146 of *Lecture Notes in Computer Science*, pages 198–218. Springer-Verlag, 1996.

[10] Roberto Segala and Nancy Lynch. Probabilistic simulations for probabilistic processes. *Nordic Journal of Computing*, 2(2):250–273, 1995. An extended abstract appeared in *Proceedings of CONCUR '94*, LNCS 836: 481-496.

[11] Paul F. Syverson and Stuart G. Stubblebine. Group principals and the formalization of anonymity. In *World Congress on Formal Methods (1)*, pages 814–833, 1999.

[12] P.F. Syverson, D.M. Goldschlag, and M.G. Reed. Anonymous connections and onion routing. In *IEEE Symposium on Security and Privacy*, pages 44–54, Oakland, California, 1997.