

Probabilistic and Information Theoretic Approaches to Anonymity

Konstantinos Chatzikokolalis a effectué sa thèse à l'Ecole Polytechnique (LIX) et à l'INRIA Saclay - Île-de-France sous la direction de Catuscia Palamidessi.



Pendant les dix dernières années, les activités sur Internet sont devenues une partie essentielle de la vie quotidienne. Comme ces activités deviennent de plus en plus fréquentes, il y a un nombre croissant d'informations personnelles qui sont enregistrées sous forme électronique et envoyées par des moyens de communication publiques. Ça rend possible – et souvent facile – de collecter et d'analyser une quantité énorme d'informations pour un individu. Par conséquent, les mécanismes pour protéger ces informations deviennent indispensables.

Un exemple récent de ce problème est le passeport “biométrique”. Ces passeports, utilisés par plusieurs pays et nécessaires pour voyager aux États Unis sans visa, portent une puce RFID contenant des informations sur le propriétaire. Cette puce peut être lue à distance, sans aucun contact avec le passeport et sans que le propriétaire

puisse savoir que son passeport transmet des informations. Il est alors nécessaire qu'un tel appareil soit muni d'un mécanisme qui assure que les informations ne seront accessibles par aucune personne non autorisée

En général, la “protection de la vie privée” (privacy) peut être définie comme la capacité de l'utilisateur à choisir et à contrôler les personnes qui auront accès à ses informations. On peut ensuite classer les propriétés de privacy en fonction de la nature des informations protégées. La *sûreté* se réfère aux informations confidentielles comme le numéro d'une carte bancaire. De l'autre côté, l'anonymat se réfère à l'identité de l'agent qui effectue une certaine action et “unlinkability” concerne le lien entre l'information et l'utilisateur. Les protocoles pour la protection de la vie privée ont comme but d'assurer cette propriété pendant une transaction électronique. Par exemple, un protocole de vote permet à l'utilisateur de voter sans révéler le lien entre le vote et son identité. Un protocole d'anonymat permet l'envoi d'un message dans un réseau public sans révéler l'identité de l'expéditeur.

Dans cette thèse, on a étudié les protocoles de protection de la vie privée, en se concentrant surtout sur l'anonymat, et on a proposé des méthodes pour exprimer et vérifier cette propriété. Les protocoles d'anonymat utilisent souvent des techniques aléatoires pour introduire du *bruit*, ce qui limite la capa-

cité d'un observateur malveillant. En plus, l'intrus peut être probabiliste lui-même, au sens qu'il pourrait utiliser des techniques d'inférence statistique pour déduire qu'un certain agent a effectué une action avec une probabilité supérieure que les autres. On a considéré un cadre probabiliste où on a un ensemble d'actions anonymes et un ensemble d'actions observables. Le comportement du protocole est spécifié par la probabilité conditionnelle d'observer chaque action observable, quand une action anonyme se produit.

Dans ce cadre, on a proposé une formalisation de l'“*innocence probable*”, une notion d'anonymat probabiliste qu'on trouve dans des protocoles “réels” tels que Crowds. On a analysé, d'une façon critique, deux définitions différentes d'innocence probable qu'on trouve dans la littérature. On a montré que notre définition généralise les deux précédentes: elle reste équivalente sous certaines hypothèses mais elle peut être appliquée dans des cas plus généraux. On a aussi étudié le comportement de cette propriété dans le cas des compositions des protocoles.

Ensuite, on a visé à une définition quantitative d'anonymat, qui nous permettrait de mesurer l'anonymat d'un protocole sur une échelle continue. En utilisant une telle définition, par exemple, on pourrait régler un paramètre du protocole en regardant comment sa modification affecte l'anonymat. On a pro-

posé une méthode pour analyser des protocoles d'anonymat en les regardant comme des "canaux" au sens de la théorie de l'information. Un canal est constitué d'un ensemble de valeurs d'entrée, un ensemble de valeurs de sortie et d'une matrice de transition qui donne la probabilité conditionnelle de produire chaque valeur de sortie pour une certaine valeur d'entrée.

Dans le cas d'anonymat, l'entrée correspond aux actions anonymes, la sortie correspond aux actions observables et la matrice de transition contient les probabilités conditionnelles qui décrivent le comportement du protocole. On peut alors appliquer des techniques de la théorie de l'information pour raisonner sur la connaissance que l'intrus peut obtenir en observant la sortie du protocole. On propose de définir le degré d'anonymat d'un protocole comme l'inverse de la *capacité* du canal correspondant. De plus, on a introduit la notion de *capacité relative* pour le cas des protocoles qui révèlent volontairement un degré limité d'information, tels que les protocoles de vote électronique. On a également étudié la relation entre cette mesure d'anonymat et celles qu'on trouve dans la littérature.

On a ensuite démontré une méthode pour analyser et vérifier un protocole: on crée d'abord un modèle du protocole en utilisant un formalisme probabiliste, tel qu'un calcul des processus probabiliste. Ensuite, on utilise un outil de modèle-checking, tel que PRISM, pour calculer la matrice du canal correspondant, et depuis cette matrice on peut calculer la capacité, soit en utilisant un algorithme général soit en exploitant certaines symétries de la matrice. On a appliqué cette technique aux protocoles «Dining Cryptographers» et «Crowds».

Par la suite, on s'est intéressé à l'intrus qui se trouve dans le scénario suivant: il ne peut pas détecter di-

rectement l'action d'intérêt, c'est à dire la valeur d'entrée, mais il peut observer la valeur de sortie, qui dépend de l'entrée selon une distribution conditionnelle connue. Ce genre de situation, qu'on trouve également dans d'autres disciplines, s'appelle "test d'hypothèse" et il a été largement étudié dans le domaine de la statistique. Une des approches les plus utilisées pour ce problème est la méthode bayésienne, qui consiste à supposer connue la distribution "a priori" des hypothèses et, à partir de là, obtenir la distribution "a posteriori", après avoir observé une certaine action.

Dans ce scénario, on a étudié l'efficacité d'une attaque potentielle en considérant la *probabilité d'erreur* de l'intrus (également appelé "*risque de Bayes*"). On a considéré la relation avec la capacité et on a montré que, dans le long terme, le risque de Bayes est indépendant de la distribution de l'entrée. On a également montré qu'un certain nombre de points, appelés "*points d'angle*", jouent un rôle important au risque de Bayes. On a obtenu une caractérisation de l'ensemble des points d'angle et on a montré qu'ils dépendent seulement de la matrice du canal. Calculer cet ensemble nous permet d'améliorer les bornes sur la probabilité d'erreur.

Ensuite on s'est concentré sur le problème de comparaison de la capacité des deux canaux, un problème difficile à cause de la complexité de la fonction de capacité. On a développé un principe de monotonie pour la capacité, basé sur sa convexité comme fonction de la matrice du canal. On a utilisé ce principe pour démontrer un nombre de résultats pour les canaux binaires. On a d'abord développé un nouvel ordre partiel pour la théorie algébrique de l'information. En utilisant le principe de monotonie, on a démontré que la capacité est monotone par rapport à cet ordre, ce qui nous permet de comparer des canaux. Ensuite, on a établi des

bornes sur la capacité basées sur des fonctions simples. On a également étudié le comportement de la capacité sur des lignes de déterminant fixe, qui mène à des méthodes graphiques pour le raisonnement sur la capacité, permettant de comparer des canaux dans la majorité des cas.

Dans l'analyse des systèmes probabilistes complexes, on veut souvent laisser une partie de leur comportement non-précisée, ce qui donne lieu à des modèles non-déterministes. Lorsque des comportements probabilistes et non-déterministes sont combinés, on introduit typiquement la notion d'"*ordonnanceur*" pour résoudre le non-déterminisme. Il a été observé que pour certaines applications, notamment celles de sécurité, l'ordonnanceur doit être restreint afin de ne pas révéler le résultat des choix probabilistes, sinon le modèle de l'intrus serait trop fort même pour des protocoles "trivialement" corrects. On a proposé un calcul des processus où le contrôle sur l'ordonnanceur peut être spécifié en termes syntaxiques, et on l'a appliqué au problème mentionné ci-dessus.

Finalement, on a proposé une variante du pi-calcul probabiliste en tant que cadre pour spécifier les protocoles de sécurité probabilistes. Afin d'exprimer et de vérifier leurs propriétés, on a développé une version probabiliste de la "sémantique de test". On a ensuite illustré cette technique sur le protocole de "Partial Secrets Exchange" qui garantit l'échange équitable d'informations entre deux agents.