EDITE de Paris
École Doctorale Informatique, Télécommunications et Électronique

Résumé en français

CEA/CESTI-LETI

Extraction de Caractéristiques pour les Attaques par Canaux Auxiliaires

Eleonora Cagli
Id. 3373691

Directeur de Thèse
Emmanuel Prouff
Encadrante
Cécile Dumas

Laboratoire d'électronique et de technologie de l'information

Commissariat á l'énergie atomique et aux énergies alternatives
MINATEC Campus | 17 rue des Martyrs | 38054 Grenoble Cedex 9
www-leti.cea.fr
Établissement public á caractère industriel et commercial RCS Paris B 775 685 019

Direction de la recherche technologique

Table des matières

1 Contexte

1.1 Le CESTI

Les présents travaux de doctorat ont été réalisés au sein du laboratoire CESTI (Centre d'Évaluation de la Sécurité des Systèmes d'Information) du CEA de Grenoble. La mission d'un CESTI est d'évaluer les aspects sécuritaires des produits qui nécessitent l'obtention d'un certificat pour pouvoir être commercialisés sur certains marchés sensibles. Les cartes à puce sont un exemple notable de tels types de dispositifs. Dans le schéma de certification français, c'est l'ANSSI (Agence National de la Sécurité des Systèmes d'Information) qui délivre le certificat, après consultation d'un rapport issu d'un des laboratoires CESTI agrées.

Un dispositif sécurisé permet, dans la grande majorité des cas, d'exécuter des algorithmes cryptographiques, pour offrir des garanties de confidentialité, authenticité, non-répudiation et intégrité des données pour les protocoles d'interface avec ce même dispositif. Quand un algorithme cryptographique est implémenté sur un support matériel, il devient potentiellement vulnérable à des attaques autres que celles considérées en cryptanalyse classique. En effet, en plus de la faiblesse mathématique théorique de l'algorithme, il existe des faiblesses matérielles liées à l'implémentation. Ces attaques matérielles sont à prendre en compte dans l'évaluation sécuritaire d'un produit sécurisé. Notamment, une partie du processus d'évaluation consiste à mener des attaques par canaux auxiliaires (ou *Side-Channel Attacks* en anglais, d'où l'acronyme SCA), qui font l'objet de cette thèse, et qui exploitent des fuites d'information issues de *canaux auxiliaires*, c'est-à-dire autres que les interfaces I/O du composant.

1.2 Les attaques par canaux auxiliaires

Introduites en 1996 par Paul Kocher [4], les attaques par canaux auxiliaires sont basées sur l'observation des variations de certaines quantités physiques du composant, comme la consommation de puissance ou le rayonnement électromagnétique, pendant l'exécution des algorithmes cryptographiques. En effet, en observant ces comportements physiques involontaires, qui sont mesurés sous forme de signaux, des déductions sur les variables internes de l'algorithme peuvent être faites. L'attaquant choisit ensuite, selon l'algorithme attaqué, les variables internes, appelées *variables sensibles*, qui seront suffisantes pour inférer la clé secrète.

2 Objectifs et contributions

Dans le contexte d'évaluation d'un dispositif sécurisé, les évaluateurs peuvent avoir accès à un ou plusieurs exemplaires du dispositif *ouverts*, ou *à secrets connus*. Ces dispositifs donnent droit à l'évaluateur de choisir ou connaître la clé secrète, de fixer d'autres variables, de désactiver des contremesures, ou de charger du logiciel. Cette possibilité est exploitée pour lancer des exécutions dans lesquelles l'attaquant aurait la connaissance complète du flux d'exécution, y compris les opérations, les variables internes manipulées, les accès aux registres, les aléas tirés en interne, etc. De cette manière il est capable de comprendre et caractériser les relations entre le comportement interne du composant et les observations physiques, avant de lancer l'attaque sur un autre dispositif, fermé, à clé inconnue. Quand une phase de caractérisation est disponible, on parle d'attaques par profilage, qui ont un rôle très important dans l'évaluation d'un dispositif, permettant de tester celui-ci dans le scénario le plus favorable pour l'attaquant. Cette thèse se concentre principalement sur cette typologie d'attaques. En effet, nous traitons les problèmes qu'un évaluateur rencontre quand, dans un scénario si favorable, il veut exploiter de façon optimale la phase de caractérisation, afin d'extraire ensuite un maximum d'information des signaux acquis dans la phase propre d'attaque. Un des premiers enjeux est la sélection des dits *points d'intérêt* (*Points of Interest* en anglais, ou PoIs), problème strictement relié au problème plus général de la réduction de dimension.

2.1 L'avant-propos de cette thèse : la recherche des points d'intérêt

L'acquisition des signaux issus des canaux auxiliaires se fait habituellement à l'aide d'un oscilloscope numérique, qui effectue un échantillonnage des signaux analogiques et les transforme en séquences numériques discrètes. Ces séquences sont souvent appelés *traces*, et leurs composantes sont les *caractéristiques* temporelles, ou points temporels, du signal. Pour garantir une analyse poussée du

dispositif, la fréquence d'échantillonnage doit être très élevée, ce qui provoque l'acquisition de traces de grand dimension. **CD : je ne pense pas que l'argument soit le plus pertinent. A discuter.** Cependant, en général, il semble qu'un nombre limité de points temporels soit nécessaire pour mener une attaque. Ce sont les PoIs, c'est-à-dire les points qui dépendent statistiquement de la variable sensible ciblée par l'attaque. En littérature l'utilisation de certains tests d'hypothèse statistique est déployée pour effectuer une sélection des PoI comme phase préliminaire d'une attaque. Cette sélection permettrait de réduire la complexité de l'attaque, en termes de temps et de mémoire. L'objectif initial de cette thèse était de proposer de nouvelles méthodes pour chercher et caractériser les PoIs, pour améliorer et possiblement optimiser ce pré-traitement des traces par sélection de points.

2.2 Approche par réduction de dimension

Au-delà de l'utilisation de statistiques univariées pour identifier les PoIs, un différent axe de recherche s'est développé dans le contexte des SCAs, issu du domaine de l'apprentissage automatique (ou *Machine Learning*, ML). Des techniques plus générales ont été proposées pour la réduction de la dimension des données, passant d'une approche par sélection de caractéristiques à une approche par *extraction de caractéristiques*. Aux alentours du 2014, les méthodes linéaires d'extraction de caractéristiques ont attiré l'attention des chercheurs, en proposant l'application de techniques telles que l'*Analyse aux Composantes Principales* (PCA), l'*Analyse Discriminante Linéaire* (LDA) ou les *Projection Pursuits* (PP). Ces méthodes exploitent des combinaisons linéaires avantageuses des points temporels, pour définir des nouvelles caractéristiques amenant à des attaques plus efficaces. La première contribution de cette thèse fait partie de cet axe de recherche : nous avons abordé deux enjeux concernants l'application des PCA et LDA dans le contexte SCA : le choix des composantes, et le problème de la taille de l'échantillonnage. Les résultats de cette étude, publiée en 2015 à CARDIS [1], sont résumés en section 3.2 et font l'objet du chapitre 4 de la thèse.

Aujourd'hui, tout dispositif sécurisé est équipé de contre-mesures spécifiques contre les SCAs. Un type de contremesure très efficace repose sur le *masquage*. Quand un masquage est implémenté correctement, toute variable interne du calcul originaire qui est sensible, est divisée en plusieurs morceaux, dont la majorité est tirée au sort pendant l'exécution. Ainsi tout sous-ensemble de morceaux est statistiquement indépendant de la variable sensible elle-même. Le calcul cryptographique est mené an accédant uniquement aux morceaux, et non pas à la variable sensible. Ceci oblige l'attaquant à analyser des distributions de probabilité conjointes des caractéristiques du signal, en étudiant conjointement son comportement aux instants temporels où chacun des morceaux est manipulé. Autrement dit, les statistiques univariées qui sont exploitables pour identifier les PoIs en absence de masquage deviennent inefficaces si un masquage est présent, car tout point temporel du signal est par lui-même indépendant de la variable sensible. En outre, les distributions jointes du signal doivent être analysées aux ordres statistiques supérieurs pour retrouver une dépendance statistique des données sensibles. Ceci implique que les méthodes linéaires d'extraction de caractéristiques sont aussi inefficaces dans ce contexte. Pour résumer, la sélection ou l'extraction de caractéristiques dans des traces protégées par masquage présente des difficultés non-négligeables. Cette complexité est mitigée quand l'attaquant peut effectuer une phase de caractérisation pendant laquelle il peut accéder aux valeurs aléatoires des parties du masquage pendant l'exécution. En pratique, ceci n'est pas tout le temps possible. Dans cette thèse on aborde ce sujet dans le cas où cette possibilité est nulle, en proposant l'exploitation de la technique de l'*Analyse Discriminante par Noyau* (*Kernel Discriminant Analysis*, KDA). Ceci est une extension de la LDA qui permet d'extraire des caractéristiques de façon non-linéaire. Les résultats obtenus dans ce contexte ont été publiés à CARDIS 2016 [2] et sont résumés en section 3.3. Ils font l'objet du chapitre 5 de la thèse.

2.3 Vers l'apprentissage profond

En observant le chemin que nous avons suivi pendant les travaux de thèse, on remarque que nous sommes partis du problème d'identifier les PoIs d'un signal, ce qui est classiquement résolu par des outils statistiques classiques, et qu'ensuite nous avons élargi à la fois les objectifs et les méthodologies. En effet, que ce qui influençait le plus la réussite d'une attaque était la qualité de l'extraction d'information. Extraire de l'information demande d'approximer des distributions de probabilité qui permettent de distinguer différent valeurs secrètes. Les premières attaques par canaux auxiliaires proposées en littérature opérait point par point, donc nécessitait d'analyser les distributions de donnée en seule-

ment quelques instants temporels pris séparement. Dans ce contexte la sélection des PoIs jouait un rôle fondamental. Cependant, dès qu'on fait un pas en arrière vers l'objectif d'une attaque, et qu'on se demande comment approximer des distributions distinguables, le fait de rejeter complètement une grande partie des caractéristiques du signal, en n'en sélectionnant que quelques unes, paraît du gaspillage. Des méthodes appropriées pour combiner ces caractéristiques peuvent mener à l'extraction de caractéristiques plus discriminantes. Pour déterminer ces combinaisons appropriées, nous avons exploré les outils d'extraction de caractéristiques afin de les utiliser comme pré-traitement du signal. En un premier temps, nous avons considéré des outils linéaires, ensuite des généralisations non-linéaires pour satisfaire une condition nécessaire pour aborder les implémentations protégées par masquage.

Conscients du fait que ces outils sont à mi-chemin entre les statistiques multivariées classiques et le domaine de l'apprentissage automatique, nous avons commencé à explorer ce domaine, qui est aujourd'hui en grand développement. Le grand intérêt pour l'apprentissage automatique est justifié par sa capacité à capter et analyser données de grande dimension dans une large variété de champs applicatifs, y compris les attaques par canaux auxiliaires. Pour cela, des modèles de plus en plus complexes sont mis en œuvre, trop complexes pour être traités dans un cadre de statistiques formelles. L'apprentissage automatique accepte des non-optimalités intrinsèques mais montre aujourd'hui d'excellents résultats.

L'étude des outils d'apprentissage automatique nous a mené à effectuer davantage un pas en arrière vers l'objectif d'une attaque : plutôt qu'optimiser des pré-traitement de données, afin d'obtenir des caractéristiques montrant des distributions facilement distinguables, nous pouvons chercher des modèles pour approximer directement ces distributions à partir des données brutes. Cette approche est propre d'une branche de l'apprentissage automatique, qui s'appelle apprentissage profond. Dans l'apprentissage profond la phase de caractérisation des données est effectuée en un seul processus, qui intègre éventuellement les pré-traitements nécessaires. Ceci est fait à l'aide de modèles multi-couches, notamment les *réseaux neuronals* (*Neural Networks*, NN), sur lesquels nous nous concentrons dans la dernière partie de la thèse. étant des modèles non-linéaires, les NN peuvent être utilisés pour adresser la contremesure de masquage. De plus, des architectures particulières de NN, les dits réseaux convolutifs (CNN), conçus originairement pour la reconnaissance d'images, s'adaptent aussi bien à d'autres types de contremesures : celles qui provoquent une désynchronisation des signaux. Nous avons étudié ce contexte, en proposant l'utilisation des CNNs comme solution, munie d'une autre stratégie classique dans le domaine de l'apprentissage automatique, l'*augmentation des données* (DA). Le chapitre 6 de la thèse est dédié à ce sujet. Les résultats obtenus ont été publiés à CHES 2017 [3] et sont résumés en section 3.4.

3 Résultats principaux

3.1 Notations

Dans la thèse, le symbole X désigne une variable aléatoire ($\vec{X}$ pour un vecteur colonne aléatoire) sur un ensemble $\mathcal{X}$, et x (respectivement $\vec{x}$) désigne une réalisation de X (respectivement $\vec{X}$). La i-ème coordonnée d'un vecteur $\vec{x}$ est indiquée par $\vec{x}[i]$, et la transposée d'un vecteur $\vec{x}$ par $\vec{x}^\intercal$. Les matrices sont indiquées par des majuscules en gras, **A** ou **S**. Les traces acquises des canaux auxiliaires sont interprêtées comme réalisations $\vec{x}_1, \ldots, \vec{x}_N$ d'un vecteur aléatoire réel $\vec{X} \in \mathbb{R}^D$, où D est la longueur du signal. Quand une méthode de réduction de dimension est utilisée comme pré-traitement, celle-ci amène à la définition d'une fonction appelée *extracteur* et dénotée par ϵ: $\mathbb{R}^D \to \mathbb{R}^C$. La variable sensible manipulée pendant l'acquisition des traces est notée Z. Celle-ci peut avoir différentes formes, mais souvent dans cette thèse elle est définie comme $Z = f(K, E)$, où E dénote une variable publique, par exemple une partie de message en clair, et K une partie d'une clé secrète que l'attaquant souhaite retrouver. Les valeurs acquises par la variable sensible sont vues comme réalisations de la variable aléatoire Z en $\mathcal{Z} = \{s_1, \ldots, s_{|\mathcal{Z}|}\}$. Les éléments de Z sont parfois encodés via le *one-hot-encoding* : à chaque élément s_j on associe un vecteur $\vec{s_j}$ de dimension $|\mathcal{Z}|$, avec toutes les entrées nulles, sauf la j-ème qui est égale à 1 : $s_j \to \vec{s_j} = (0, \ldots, 0, \underbrace{1}_{j}, 0, \ldots, 0)$. Un élément générique de

$\mathcal{Z}$ sera noté s, si son indice i n'est pas nécessaire.

3.2 Techniques Linéaires de Réduction de Dimension

Dans cette section sont décrites les études menées autour des méthodes linéaires d'extraction de caractéristiques, en particulier de l'Analyse aux Composantes Principales (PCA) et de l'Analyse Discriminante Linéaire (LDA).

3.2.1 Analyse aux Composantes Principales, l'outil classique et le profilée

L'extracteur linéaire $\epsilon^{\mathrm{PCA}}(\vec{x}) = \mathbf{A}\vec{x}$ se déduit des certains vecteurs propres $\vec{\alpha}_1, \ldots, \vec{\alpha}_C$, appelés *Composantes Principales* (PCs), dont les transposés sont arrangé en tant que lignes dans la matrice de projection **A**. Classiquement la PCA intervient sur des données non labellisées $\vec{x}_1, \ldots, \vec{x}_N$, supposé ayant moyenne nulle et arrangé comme colonnes dans une matrice **M** de dimension $D \times N$, de tel sort que la matrice de covariance des données est la suivante :

$$\mathbf{S} = \frac{1}{N}\mathbf{M}\mathbf{M}^{\intercal} \ . \tag{1}$$

Dans ce cas, les vecteur propres $\vec{\alpha}_1, \ldots, \vec{\alpha}_C$ correspondent aux vecteurs propres de la matrice **S** et leurs valeurs propres associées sont notées $\lambda_1, \ldots, \lambda_r$. La PCA est la projection qui maximise la variance globale des caractéristiques extraites. La variance étant liée à la quantité d'information des données, cette transformation est censée réduire la dimension des traces tout en renforçant l'information contenue. Une propriété remarquable de la PCA est que chaque λ_i correspond à la variance empirique des données projetées sur la PC correspondante $\vec{\alpha}_i$.

Dans un scénario d'attaque profilée, cet outil classique est toutefois largement sous-optimal : il n'exploite pas une phase de caractérisation. Dans cette dernière on suppose que l'attaquant est en possession d'un ensemble de données étiquet'ees $(\vec{x}_i, z_i)_{i=1..N_p}$, c'est-à-dire où l'association *trace-variable sensible* est connue. Dans la littérature SCA [**?**, **?**, **?**, **?**, **?**] une version *profilée* de la PCA a été introduite. En introduisant les moyennes empiriques par classe

$$\vec{\mu}_s = \hat{\mathbb{E}}[\vec{X} \mid Z = s] = \frac{1}{N_s} \sum_{i:\ z_i = s} \vec{x}_i \ , \tag{2}$$

la PCA profilée utilise la matrice des *écarts inter-classes* suivante à la place de la matrice de covariance **S** :

$$\mathbf{S_B} = \sum_{s \in \mathcal{Z}} N_s (\vec{\mu}_s - \overline{\vec{x}})(\vec{\mu}_s - \overline{\vec{x}})^{\intercal} \ , \tag{3}$$

où $\overline{\vec{x}}$ est la moyenne empirique de toutes les données confondues. L'extracteur obtenu de cette manière garantie que les centroïdes par classes des données projetées sont écartés au maximum.

3.2.2 LDA and the Small Sample Size problem

3.3 Analyse Discriminante par Noyau

3.4 Réseau Neuronal Convolutif

4 Conclusions et Perspectives

Références

[1] Eleonora Cagli, Cécile Dumas, and Emmanuel Prouff. Enhancing dimensionality reduction methods for side-channel attacks. In *International Conference on Smart Card Research and Advanced Applications*, pages 15–33. Springer, 2015.

[2] Eleonora Cagli, Cécile Dumas, and Emmanuel Prouff. Kernel discriminant analysis for information extraction in the presence of masking. In *International Conference on Smart Card Research and Advanced Applications*, pages 1–22. Springer, 2016.

[3] Eleonora Cagli, Cécile Dumas, and Emmanuel Prouff. Convolutional neural networks with data augmentation against jitter-based countermeasures - profiling attacks without pre-processing. In Wieland Fischer and Naofumi Homma, editors, *Cryptographic Hardware and Embedded Systems - CHES 2017 - 19th International Conference, Taipei, Taiwan, September 25-28, 2017, Proceedings*, volume 10529 of *Lecture Notes in Computer Science*, pages 45–68. Springer, 2017.

[4] Paul C Kocher. Timing attacks on implementations of diffie-hellman, rsa, dss, and other systems. In *Annual International Cryptology Conference*, pages 104–113. Springer, 1996.