Proposition de sujet de thèse

Logique outillée pour le raffinement et la vérification de propriétés pour améliorer le respect de la vie privée

CEA LIST, Laboratoires LECS et LSL

June 7, 2018

1 Cadre

Le CEA LIST est un centre de recherche technologique sur les systèmes à logiciel prépondérant qui mène ses recherches en partenariat avec les grands acteurs industriels de l'automobile, de l'aéronautique, de la défense et du médical pour étudier et développer des solutions innovantes adaptées à leurs besoins. Au sein du CEA LIST, le Laboratoire d'Exigences et Conformité de Systèmes (LECS) et le Laboratoire de Sûreté et de Sécurité des Logiciels (LSL), localisés à Saclay (Essonne, 91), développent des outils d'aide à la modélisation, la validation par tests et la vérification de logiciels et de systèmes matériels/logiciels. Parmi ces outils, on peut citer Papyrus (https://www.eclipse.org/papyrus/) pour la modélisation des exigences et systèmes, DIVERSITY (https://projects.eclipse.org/proposals/eclipse-formal-modeling-project) supportant l'exécution symbolique des tests, et Frama-C (https://frama-c.com/), facilitant la vérification des propriétés de programmes C. La thèse se déroulera au sein du LECS et sera encadrée par des chercheurs de ces deux laboratoires.

2 Contexte et problématique

Les systèmes numériques manipulent et stockent des volumes de données de plus en plus importants, notamment des données sensibles concernant des informations personnelles et confidentielles sur leurs utilisateurs. La collecte et l'utilisation de ces données, à l'insu des utilisateurs, peuvent être exploitées à des fins diverses et avoir des impacts négatifs importants sur le respect de leur vie privée, comme la divulgation d'informations personnelles (concernant par exemple les orientations politiques, religieuses ou sexuelles des utilisateurs) ou des sollicitations non désirées (par exemple, du démarchage publicitaire agressif).

Pour améliorer le respect de la vie privée des personnes, des réglementations visent à leur donner les moyens de leur "auto-détermination informationnelle"[6] à travers des mesures de Protection des Données Personnelles (PDP). Ainsi, le RGPD[9] formule le besoin d'intégrer des exigences de PDP dans l'ingénierie des systèmes. Pour cela, le principe de "*Data Protection by Design*" stipule que les aspects de PDP doivent être pris en compte dès le début de la conception du système. Ces aspects peuvent être exprimés sous la forme de propriétés. Dans cette démarche, une des étapes essentielles consiste à formaliser et prouver que le système cible satisfait ces propriétés. Néanmoins, la majeure partie des méthodes d'analyse existantes est, tout comme les cadres normatifs actuels, rédigée en langage naturel, ce qui limite leur portée et la validité des résultats obtenus.

Des cadres formels pour vérifier des propriétés de sécurité comme la confidentialité ou l'authentification existent néanmoins[1, 4]. Ils ne sont cependant pas particulièrement adaptés pour représenter et vérifier des propriétés de PDP telles que la pseudonymisation, l'anonymisation, la chaînabilité (*linkability*), et d'autres sur la confidentialité et l'observabilité des données requérant plusieurs niveaux de granularité. De nouvelles approches sont ainsi nécessaires pour formaliser et supporter la vérification de ces propriétés de PDP.

3 Objectifs de la thèse

L'objectif général de la thèse est le développement d'une logique outillée permettant la représentation des systèmes avec flot des données et la vérification de propriétés de PDP. Pour cela, il convient de spécifier un langage formel pour représenter des systèmes et processus incluant notamment des flots de données, des utilisateurs et des unités de traitement et de stockage. Le langage doit permettre la vérification d'exigences de haut niveau utilisées par des méthodes comme LINDDUN [3] (*Linkability, Identifiability, Non-repudiation, Detectability, Disclosure of*

information, Unawareness, Non-compliance). Ce langage doit être doté d'une sémantique formelle (par exemple, une sémantique opérationnelle) afin de pouvoir analyser les aspects dynamiques du système (ses comportements).

Une fois ce langage établi, des algorithmes devront être conçus et implémentés afin de pouvoir vérifier des propriétés et des relations associées à des exigences de haut niveau. Pour cela, il faudra identifier les (sous-)propriétés pouvant être vérifiées au niveau du modèle et celles pouvant l'être au niveau du code exécutable (par exemple, au niveau d'un programme C). Une approche par raffinement, ainsi que l'utilisation d'outils de vérification existants (comme ceux de la plateforme Frama-C [5]) pourront être considérées.

4 Hypothèses de recherche

L'état de l'art propose des langages pour aborder les aspects "*Privacy by Design*", par exemple [8]. Par ailleurs, l'utilisation d'approches comme la théorie de l'information [2] est une piste à explorer afin d'avoir différents niveaux de granularité pour la représentation des données. D'autres théories comme les systèmes logiques et plus concrètement la logique épistémique dynamique [7] pourraient également fournir des bases pour la définition d'une sémantique et d'algorithmes de vérification.

5 Informations générales

Profil candidat: maîtriser au moins un des domaines suivants est nécessaire pour cette thèse: *privacy by design*, langage et sémantique formels, validation de modèles, vérification de programmes

Encadrement: Thibaud Antignac, Gabriel Pedroza, Julien Signoles, Christophe Gaston
(*prenom.nom@cea.fr*)

Processus administratif: délais administratifs de 2 ou 3 mois à prévoir.

References

[1] B. Blanchet. Automatic Verification of Correspondences for Security Protocols. *J. Comput. Secur.*, 17(4):363–434, dec 2009.

[2] C. E. Shannon. A Mathematical Theory of Communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, jan 2001.

[3] DistriNet Research Group, KU Leuven. LINDDUN: a privacy threat analysis framework. 2017.

[4] J. Stern. Why Provable Security Matters? In *Proceedings of the 22Nd International Conference on Theory and Applications of Cryptographic Techniques*, EUROCRYPT'03, pages 449–461, Berlin, Heidelberg, 2003. Springer-Verlag.

[5] F. Kirchner, N. Kosmatov, V. Prevosto, J. Signoles, and B. Yakobowski. Frama-C: A Software Analysis Perspective. *Formal Aspects of Computing*, pages 1–37, January 2015.

[6] Antoinette Rouvroy and Yves Poullet. The right to informational self-determination and the value of self-development: Reassessing the importance of privacy for democracy. In Serge Gutwirth, Yves Poullet, Paul De Hert, Cécile de Terwangne, and Sjaak Nouwt, editors, *Reinventing Data Protection?*, pages 45–76, Dordrecht, 2009. Springer Netherlands.

[7] S. Frittella and G. Greco and A. Kurz and A. Palmigiano and V. Sikimić. A proof-theoretic semantic analysis of dynamic epistemic logic. *Journal of Logic and Computation*, 26(6):1961–2015, 2016.

[8] T. Antignac and R. Scandariato and G. Schneider. Privacy Compliance via Model Transformations. In *International Workshop on Privacy Engineering (IWPE'18)*, IEEE proceedings, London, UK, 27 April 2018. To appear.

[9] The European Commission. *General Data Protection Regulation (GDPR)*. Official Journal of the European Commission, 2016.