Reconstruction et analyse sémantique de chronologies cybercriminelles

Yoan Chabot, Aurélie Bertaux, Tahar Kechadi, Christophe Nicolle

Reconstruction et analyse sémantique de chronologies cybercriminelles

Yoan Chabot[*,**] Aurélie Bertaux[**]
Tahar Kechadi[*], Christophe Nicolle[**]

[*]School of Computer Science and Informatics, University College Dublin, Ireland
yoan.chabot@ucdconnect.ie
[**]Equipe CheckSem, Laboratoire Le2i, UMR CNRS 6306,
Faculté des sciences Mirande, 21078 Dijon, France
http://checksem.u-bourgogne.fr

Résumé. La reconstruction de chronologies d'évènements cybercriminels (ou reconstruction d'évènements) est une étape primordiale dans une investigation numérique. Cette phase permet aux enquêteurs d'avoir une vue des évènements survenus durant un incident. La reconstruction d'évènements requiert l'étude d'importants volumes de données en raison de l'omniprésence des nouvelles technologies dans notre quotidien. De plus, les conclusions produites se doivent de respecter les critères fixés par la justice. Afin de répondre à ces challenges, nous proposons une nouvelle méthodologie basée sur une ontologie permettant d'assister les enquêteurs tout au long du processus d'enquête.

1 Introduction

En raison de l'évolution des nouvelles technologies, le domaine de la criminalistique informatique se heurte à des problèmes qui étaient encore anecdotiques il y a quelques années. Bien que des outils existent pour aider les enquêteurs, leur portée est limitée aux premières étapes du processus d'investigation défini par (Palmer, 2001). La collecte et l'étude des caractéristiques des pièces à conviction sont d'importantes phases du processus, toutefois il est également nécessaire de déduire de nouvelles connaissances telles que les raisons de l'état actuel des pièces à conviction (Carrier et Spafford, 2004) pour produire des conclusions utiles dans un procès. La reconstruction d'évènements peut être vue comme un processus utilisant un ensemble de pièces à conviction pour produire une chronologie décrivant les évènements composant un incident. Dans ce papier, nous présentons l'approche SADFC (Semantic Analysis of Digital Forensic Cases) qui permet de reconstruire et d'analyser des chronologies à partir de sources de données hétérogènes (traces laissées sur une scène de crime). La section suivante passe en revue les approches de reconstruction existantes. La section 3 présente ensuite notre approche et expose notamment les aspects relatifs à la gestion des connaissances et aux possibilités de raisonnements offertes. La section 4 introduit un prototype ayant permis la validation expérimentale de notre approche. Enfin, les travaux futurs sont présentés dans la section 5.

2 Étude des approches de reconstruction d'évènements

Volume et hétérogénéité des données La taille des données à traiter et leur hétérogénéité (due à l'utilisation de nombreuses sources de traces numériques telles que les fichiers de journalisation, les historiques de navigateur Web...) sont deux challenges majeurs de la reconstruction d'évènements. Dans une large part des approches existantes, des solutions sont proposées pour l'extraction automatique des évènements à partir de sources hétérogènes et la construction de la chronologie. Pour cela, des extracteurs automatiques et dédiés à chaque source d'évènements sont utilisés pour peupler un élément de stockage central (base de données (Chen et al., 2003), ontologie (Schatz et al., 2004), etc.). Concernant l'analyse de scénarios, les approches existantes proposent des fonctionnalités permettant de corréler des évènements (Schatz et al., 2004) ou d'aider l'enquêteur dans la lecture de la chronologie en produisant des évènements de haut niveau conceptuel à partir d'évènements extraits depuis des sources de traces (Hargreaves et Patterson, 2012). Toutefois, aucune des approches étudiées ne proposent une solution complète pour assister les enquêteurs dans l'interprétation et l'analyse des chronologies.

Exigences légales Les approches de reconstruction d'évènements doivent également satisfaire un ensemble d'exigences telles que la crédibilité des conclusions produites, l'intégrité des données utilisées et la reproductibilité du processus d'investigation (Baryamureeba et Tushabe, 2004). De plus, (Gladyshev et Patel, 2004) avance qu'une formalisation du problème de reconstruction d'évènements est nécessaire afin de mieux structurer le processus de reconstruction, de faciliter son automatisation et d'assurer la complétude de la reconstruction. Un problème récurrent parmi les approches étudiées est le manque de fondements théoriques permettant de valider et d'expliquer les conclusions produites.

3 Représentation de connaissances et analyse de chronologies cybercriminelles

Pour répondre aux limites précédentes, nous présentons l'approche SADFC qui permet d'assister les enquêteurs depuis l'extraction des traces numériques jusqu'à l'interprétation de la chronologie. Cette approche s'appuie sur l'analyse avancée de chronologies cybercriminelles basée sur une représentation des connaissances décrivant les activités d'un utilisateur sur un ordinateur. Bien que les informations temporelles des évènements soient une dimension primordiale, d'autres aspects doivent également être pris en compte pour offrir des fonctions d'analyse avancées. L'approche SADFC introduit une nouvelle représentation sémantiquement riche des évènements et de leurs interactions avec l'environnement intégrant les notions de *scène de crime* (espace virtuel où se déroule un ensemble d'évènements illicites), *d'évènements* (action survenant à un instant donné), *d'incidents* (ensemble des évènements illicites et d'évènements corrélés à ces derniers), *de traces* (résidus laissés par un évènement et permettant sa reconstruction), *d'objets* (ressources utilisées, générées, modifiées ou supprimées par les évènements) et *de sujets* (processus ou personnes initiant ou subissant les évènements). Le modèle proposé définit également les relations entre ces concepts parmi lesquelles les relations de *composition* (liant un évènement avec les évènements le composant), de *participation* (liant un sujet aux évènements auxquels il prend part), d'*utilisation* (liant un évènement aux objets

qu'il utilise) ou encore de corrélation (liant deux évènements interdépendants).
Pour aider les enquêteurs à mener à bien leurs enquêtes, des opérateurs de construction de chronologies et d'analyse sont proposés. Trois ensembles d'opérateurs sont définis dans l'approche SADFC :

- Les opérateurs *d'extraction* ayant pour fonction d'extraire les informations pertinentes contenues dans les traces numériques et de peupler la base de connaissances en conséquence.
- Les opérateurs *d'inférence* permettant d'enrichir la base avec de nouvelles connaissances déduites à partir des connaissances existantes. Par exemple, la seule information disponible pour déterminer le sujet impliqué dans un évènement d'un navigateur Web est son identifiant de session, présent dans certaines traces numériques produites par les navigateurs. Pour identifier le sujet impliqué dans d'autres actions, nous utilisons un opérateur d'inférence basé sur l'hypothèse suivante : soit e_i la première visite d'une page Web d'une session s, e_j la dernière visite de cette même session, t_i la date de début de e_i et t_j la date de fin de e_j, un évènement survenant sur la machine à une date comprise dans l'intervalle de temps défini par t_i et t_j implique la personne identifiée par la session s.
- Les opérateurs *d'analyse* utilisés pour aider les enquêteurs dans l'interprétation des informations portées par une chronologie.

Les opérateurs proposés dans cette section permettent de dispenser les enquêteurs des tâches les plus fastidieuses de la reconstruction d'évènements, leur permettant ainsi de se concentrer sur des tâches où leur expertise et leur expérience sont les plus utiles.

4 Implémentation

Pour valider la pertinence du modèle et la viabilité des opérateurs proposés, un prototype a été développé. Ce dernier est centré sur une ontologie (permettant notamment l'utilisation de processus automatiques pour raisonner sur les connaissances) OWL implémentant le modèle présenté dans la section précédente. L'ontologie proposée est divisée en trois couches. La couche *Provenance Knowledge Layer* contient des informations sur la manière dont l'investigation est menée (actions entreprises par les enquêteurs, informations utilisées pour parvenir à une conclusion, etc.). La couche *Common Knowledge Layer* contient des connaissances génériques sur les évènements telles que des informations temporelles, les ressources ou encore les personnes et les processus participant à leur exécution. La couche *Specialized Knowledge Layer* contient les caractéristiques spécifiques portées par chaque évènement. La modélisation de connaissances spécialisées au sein de l'ontologie permet de bénéficier de l'expertise des acteurs du domaine durant l'analyse de la chronologie.
Le prototype propose également une implémentation des trois ensembles d'opérateurs au sein d'une application Java. Les opérateurs d'extraction prennent la forme d'un ensemble d'extracteurs et de ponts permettant de peupler l'ontologie à partir de sources de traces. Les opérateurs d'inférence sont implémentés à l'aide de requêtes SPARQL/Update recherchant des motifs particuliers dans l'ontologie et ajoutant des connaissances en conséquence. Enfin, le prototype implémente un opérateur permettant d'identifier des couples d'évènements potentiellement corrélés en se basant sur des critères tels que la proximité temporelle, l'utilisation de ressources communes, les sujets participants ou encore des règles formulées par les experts du domaine. Par exemple, soit un évènement représentant la visite d'une page web et un

évènement de création de marque-page pour cette même page Web, ces deux évènements sont corrélés car ils utilisent une même ressource (la page Web) et sont créés par le même processus (le navigateur Web Firefox par exemple).

5 Conclusion et travaux futurs

Dans cet article, nous avons présenté l'approche SADFC permettant d'aider les enquêteurs durant la reconstruction et l'analyse de chronologies dans le respect des contraintes juridiques. Notre principale contribution est l'introduction d'un nouveau modèle pour décrire des incidents cybercriminels et d'opérateurs permettant le peuplement de l'ontologie ainsi que l'analyse des connaissances. L'implémentation de ces éléments au sein d'un prototype a permis de valider expérimentalement la faisabilité et la pertinence de l'approche proposée. Les travaux futurs s'intéressent à l'intégration de nouvelles sources de traces, à l'enrichissement de l'ontologie et à la conception de nouveaux opérateurs d'analyse.

Références

Baryamureeba, V. et F. Tushabe (2004). The enhanced digital investigation process model. In *Proceedings of the Fourth Digital Forensic Research Workshop*. Citeseer.

Carrier, B. et E. Spafford (2004). Defining event reconstruction of digital crime scenes. *Journal of forensic sciences 49*(6), 1291–1298.

Chen, K., A. Clark, O. De Vel, et G. Mohay (2003). Ecf-event correlation for forensics. In *First Australian Computer Network and Information Forensics Conference*, Perth, Australia, pp. 1–10. Edith Cowan University.

Gladyshev, P. et A. Patel (2004). Finite state machine approach to digital event reconstruction. *Digital Investigation 1*(2), 130–149.

Hargreaves, C. et J. Patterson (2012). An automated timeline reconstruction approach for digital forensic investigations. *Digital Investigation 9*, 69–79.

Palmer, G. (2001). A road map for digital forensic research. In *First Digital Forensic Research Workshop, Utica, New York*, pp. 27–30.

Schatz, B., G. Mohay, et A. Clark (2004). Rich event representation for computer forensics'. *Proceedings of the Fifth Asia-Pacific Industrial Engineering and Management Systems Conference (APIEMS 2004) 2*(12), 1–16.

Summary

Event reconstruction is one of the most important steps in digital forensic investigations. It allows investigators to have a clear view of the events occurring over time. Event reconstruction is a complex task which requires the exploration of a large amount of events due to the pervasiveness of new technologies nowadays. Any evidence produced must also meet the requirements of the courts. For this purpose, we propose a new approach, based on an ontology, able to assist investigators through the whole investigative process.