Reconstruction et analyse sémantique de chronologies cybercriminelles

CONFERENCE PAPER · JANUARY 2014

4 AUTHORS:

Yoan Chabot
University of Burgundy
7 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Aurélie Bertaux
University of Burgundy
23 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Tahar Kechadi
University College Dublin
214 PUBLICATIONS **682** CITATIONS
SEE PROFILE

Christophe Nicolle
University of Burgundy
96 PUBLICATIONS **141** CITATIONS
SEE PROFILE

Reconstruction et analyse sémantique de chronologies cybercriminelles

Résumé. La reconstruction de scénarios est l'une des étapes les plus importantes d'une investigation numérique. Elle permet aux enquêteurs d'avoir une vue des évènements survenus durant un incident. La reconstruction de scénarios est une tâche complexe requérant l'étude d'un très grand nombre d'évènements 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 intégrant les connaissances d'experts des domaines de la criminalistique et de l'ingénierie logicielle, permettant d'assister les enquêteurs tout au long du processus d'enquête.

1 Introduction

En raison de l'évolution des nouvelles technologies et de leur omniprésence dans notre quotidien, le domaine de la criminalistique informatique se heurte aujourd'hui à de nouveaux problèmes, encore anecdotiques il y a quelques années. Bien que des outils d'investigation numérique existent pour aider les investigateurs durant une enquête, leur portée est limitée aux premières étapes du processus d'investigation défini par (Palmer, 2001), à savoir la préservation de l'état des systèmes numériques et leur examen à la recherche de pièces à conviction. Bien que la collecte des pièces et l'étude de leurs caractéristiques soient d'importantes phases dans le processus, 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 de scénarios (dont le but est de répondre à des questions comme "Que s'est-il passé ?", "Qui est responsable ?" et "Pourquoi ces évènements ont-ils eu lieu ?") 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 répond à deux objectifs : la création d'un système capable de reconstruire des scénarios à partir de sources de données hétérogènes (traces laissées sur une scène de crime)

et de les analyser à l'aide d'outils sémantiques basés sur les connaissances d'experts du domaine pour faciliter la prise de décisions et la production de conclusions par les enquêteurs. Par ailleurs, notre architecture répond à des exigences, en terme de qualité du processus d'investigation, pour satisfaire les attentes de la justice. La section suivante passe en revue les défis inhérents à la reconstruction de scénarios et les solutions proposées par les approches existantes. La section 3 présente ensuite les aspects de l'approche SADFC relatifs à la gestion des connaissances. Un aperçu des possibilités de raisonnements offertes est ensuite donné dans la section 4. Enfin, les travaux futurs sont présentés dans la section 5.

2 Étude des approches de reconstruction de scénarios

La reconstruction de scénarios présente plusieurs difficultés directement liées à la taille et à la nature des volumes de données à traiter ainsi qu'à la complexité des processus juridiques. Le tableau 1 compare les forces (✓) et les faiblesses (absence de solution (✖) ou solution partielle ou inadaptée (●)) de plusieurs approches au regard des défis les plus importants du domaine.
Bien que le problème de reconstruction de scénarios ait été l'objet de nombreux travaux durant la dernière décennie, 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...) restent problématiques à l'heure actuelle. La capacité des approches existantes à répondre à ces défis est évaluée à l'aide des trois critères suivants :

- La mise en place d'outils permettant d'automatiser l'extraction d'évènements et d'ordonner automatiquement ces évènements dans une chronologie (voir critère (1) du tableau 1).
- La capacité de l'outil proposé à pouvoir traiter des sources d'évènements hétérogènes et à fédérer les informations obtenues dans un modèle de manière cohérente et structurée pour faciliter l'analyse des données (voir critère (2) du tableau 1).
- La capacité des outils proposés à assister l'enquêteur dans les tâches d'analyse de la chronologie (voir critère (3) du tableau 1).

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. Dans la proposition de l'architecture ECF, (Chen et al., 2003) introduisent l'utilisation d'un ensemble d'extracteurs automatisés et dédiés à chaque source pour collecter des évènements et stocker ces derniers dans un élément de stockage unique constituant la chronologie. Cette idée s'est largement répandue dans la littérature (Olsson et Boldt, 2009), (Gudhjonsson, 2010), (Hargreaves et Patterson, 2012). Par ailleurs, le système FORE proposé par (Schatz et al., 2004) introduit le stockage par une ontologie pour bénéficier d'une représentation sémantiquement riche des évènements. Concernant l'analyse de scénarios, peu de solutions sont proposées dans la littérature. L'approche FORE propose un système permettant d'identifier des corrélations entre évènements en les connectant avec des liens de causes et d'effets. (Gladyshev et Patel, 2004) tentent de mener à bien la reconstruction de scénarios en représentant le comportement du système sujet de l'investigation à l'aide d'automates à états finis puis en recherchant des séquences satisfaisant les contraintes imposées par les pièces à conviction (le processus décrit n'est toutefois pas automatisé). (Hargreaves et Patterson, 2012) utilise un processus basé sur des patrons pour produire des évènements de haut niveau conceptuel (facilement compréhensible par l'humain) à partir d'une chronologie contenant des évènements de

Approche / Critère	(1)	(2)	(3)	(4)	(5)
ECF (Chen et al., 2003)	✓	✓	✖	✖	✖
FORE (Schatz et al., 2004)	✓	✓	●	✖	✖
Finite state machine(Gladyshev et Patel, 2004)	✖	●	●	✓	✖
Zeitline (Buchholz et Falk, 2005)	●	✓	✖	✖	✓
Neural networks(Khan et Wakeman, 2006)	●	●	✖	●	✖
CyberForensic TimeLab(Olsson et Boldt, 2009)	✓	✓	✖	✖	✖
log2timeline (Gudhjonsson, 2010)	✓	✓	✖	✖	✖
Timeline reconstruction (Hargreaves et Patterson, 2012)	✓	✓	●	●	✖

TAB. 1 – *Comparaison des approches*

bas niveau conceptuel. Bien que cette approche soit pertinente, elle permet seulement de gérer un des nombreux aspects de l'analyse en proposant aux enquêteurs un résumé de la chronologie. Les autres aspects tels que la corrélation d'évènements ne sont notamment pas couverts par cette approche.
Les approches de reconstruction de scénarios 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). Pour étudier la capacité des approches existantes à répondre à ces exigences, les critères suivants sont ajoutés à notre étude :

- La présence d'une théorie pour étayer l'approche proposée et la capacité à expliciter les raisonnements effectués par l'outil (voir critère (4) du tableau 1).
- La capacité à préserver l'intégrité des informations (voir critère (5) du tableau 1).

En préambule de leurs travaux, (Gladyshev et Patel, 2004) avancent qu'une formalisation du problème de reconstruction de scénarios est nécessaire afin de mieux structurer le processus de reconstruction, de faciliter son automatisation et d'assurer la complétude de la reconstruction. (Khan et Wakeman, 2006) proposent un système de reconstruction de scénarios basés sur les réseaux de neurones. Toutefois, bien qu'étant une théorie reconnue, les réseaux de neurones ne permettent pas d'expliciter suffisamment les raisonnements utilisés. Enfin, (Hargreaves et Patterson, 2012) conservent des informations sur le raisonnement ayant permis d'inférer chaque évènement de haut niveau offrant ainsi aux enquêteurs la possibilité de fournir de plus amples informations à la justice. Concernant la préservation de l'intégrité des informations (Buchholz et Falk, 2005) définissent un ensemble de restrictions pour éviter l'altération des pièces à conviction par l'outil.

En conclusion, deux principales limitations des approches existantes sont le manque d'outils pour analyser de manière automatique les chronologies produites et le manque de fondements théoriques permettant de valider et d'expliquer les conclusions produites. C'est pourquoi nous présentons l'approche SADFC qui permet d'assister les enquêteurs depuis l'extraction des traces numériques jusqu'à la construction de la chronologie et son interprétation. 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.

3 Gestion et représentation de connaissances décrivant des incidents cybercriminels

La proposition d'une nouvelle représentation pour les connaissances liées à une affaire cybercriminelle est motivée par les capacités limitées des approches existantes en terme d'analyse. Bien que les informations temporelles des évènements soient une dimension primordiale, d'autres aspects tels que les liens de causalité entre évènements ou encore les ressources utilisées par ces derniers doivent également être pris en compte pour offrir des fonctions d'analyse avancées. Nous proposons donc 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'incident* (ensemble des évènements illicites et d'évènements corrélés à ces derniers), *de traces* (résidu laissé 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 (e.g. causalité)).

L'approche SADFC propose une implémentation de ce modèle au sein d'une ontologie. Le recours à cette dernière est motivée par leur pertinence démontrée dans les travaux de (Schatz et al., 2004). D'après (Gruber et al., 1993), « une ontologie est une spécification explicite et formelle d'une conceptualisation partagée ». La nature formelle et explicite de l'ontologie permet l'utilisation de processus automatiques pour raisonner sur les connaissances. De plus, l'ontologie permet de disposer d'une sémantique riche pour représenter les connaissances, selon une vision du domaine de la criminalistique partagée par les experts du domaine.

L'ontologie proposée dans l'approche SADFC est divisée en trois couches :

- La couche PKL (*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.) et permet ainsi de satisfaire les exigences de la justice.
- La couche CKL (*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. Cette couche offre une représentation pour les évènements facilitant ainsi les raisonnements sur les connaissances.
- La couche SKL (*Specialized Knowledge Layer*) contient les caractéristiques spécifiques portées par chaque évènement. L'intégration de connaissances spécialisées est un processus complexe nécessitant le concours d'experts du domaine. La modélisation de ces connaissances au sein de l'ontologie permet de bénéficier de leur expertise durant l'analyse de la chronologie.

L'introduction d'une représentation des évènements à l'aide de nombreuses dimensions et l'implémentation de ce modèle à l'aide d'une ontologie permet d'augmenter la capacité d'analyse de notre approche. Pour traiter les connaissances de l'ontologie et aider les enquêteurs à me-

ner à bien leurs enquêtes, des opérateurs de construction de chronologies et d'analyse sont proposés.

4 Opérateurs de construction et d'analyse avancée de chronologies cybercriminelles

Trois ensemble d'opérateurs sont définis dans l'approche SADFC :

- Les opérateurs *d'extraction* ayant pour fonction d'identifier et d'extraire les informations pertinentes contenues dans les traces numériques issues de diverses sources (dans nos travaux, nous travaillons notamment à partir d'historiques de navigateurs Web). Ces opérateurs créent ensuite les évènements associés à ces traces et peuplent l'ontologie en conséquence.
- Les opérateurs *d'inférence* permettant d'enrichir l'ontologie 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. Nous proposons notamment 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.

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

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 de représentation des connaissances pour décrire des incidents cybercriminels. De plus, des opérateurs permettant le peuplement de l'ontologie ainsi que l'analyse de ces connaissances sont également proposés. Les premiers résultats obtenus ont montré que l'utilisation d'une telle représentation des évènements permet de réaliser des tâches d'analyse avancées telles que l'identification de corrélations entre évènements.

Les travaux futurs sur l'approche SADFC s'intéressent à la conception d'une architecture logicielle centrée sur l'ontologie développée afin de démontrer expérimentalement la pertinence de l'approche SADFC.

Références

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

Buchholz, F. et C. Falk (2005). Design and implementation of zeitline : a forensic timeline editor. In *Digital forensic research workshop*.

Carrier, B. D. et E. H. Spafford (2004). Defining event reconstruction of digital crime scenes. *Journal of Forensic Sciences 49*(6), 1291.

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.

Gruber, T. R. et al. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition 5*(2), 199–220.

Gudhjonsson, K. (2010). Mastering the super timeline with log2timeline. *SANS Reading Room*.

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

Khan, M. et I. Wakeman (2006). Machine learning for post-event timeline reconstruction. In *First Conference on Advances in Computer Security and Forensics Liverpool, UK*, pp. 112–121.

Olsson, J. et M. Boldt (2009). Computer forensic timeline visualization tool. *Digital Investigation 6*, 78–87.

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 occuring over time. Event reconstruction is a complex task which requires 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 which can assist investigators through the whole investigative process. The proposed approach is based on an ontology which integrates knowledge of experts from the fields of digital forensics and software development and allows a semantically rich representation of events related to the incident. The main purpose of this ontology is to analyse these events in an automatic way.