Digital Investigation 24 (2018) S2–S9

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

DFRWS 2018 Europe — Proceedings of the Fifth Annual DFRWS Europe

Using computed similarity of distinctive digital traces to evaluate non-obvious links and repetitions in cyber-investigations

Timothy Bollé*, Eoghan Casey

University of Lausanne, School of Criminal Justice, 1015, Lausanne-Dorigny, Switzerland

Keywords:
Digital forensics
Digital traces
Digital evidence
Similarity measures
Email similarity
Trace similarity
Case comparisons
Case linkage
Cyber-investigation
Near similarity computation
Crime analysis
Forensic intelligence

A B S T R A C T

This work addresses the challenge of discerning non-exact or non-obvious similarities between cybercrimes, proposing a new approach to finding linkages and repetitions across cases in a cyber-investigation context using near similarity calculation of distinctive digital traces. A prototype system was developed to test the proposed approach, and the system was evaluated using digital traces collected during actual cyber-investigations. The prototype system also links cases on the basis of exact similarity between technical characteristics. This work found that the introduction of near similarity helps to confirm already existing links, and exposes additional linkages between cases. Automatic detection of near similarities across cybercrimes gives digital investigators a better understanding of the criminal context and the actual phenomenon, and can reveal a series of related offenses. Using case data from 207 cyber-investigations, this study evaluated the effectiveness of computing similarity between cases by applying string similarity algorithms to email addresses. The Levenshtein algorithm was selected as the best algorithm to segregate similar email addresses from non-similar ones. This work can be extended to other digital traces common in cybercrimes such as URLs and domain names. In addition to finding linkages between related cybercrime at a technical level, similarities in patterns across cases provided insights at a behavioral level such as modus operandi (MO). This work also addresses the step that comes after the similarity computation, which is the linkage verification and the hypothesis formation. For forensic purposes, it is necessary to confirm that a near match with the similarity algorithm actually corresponds to a real relation between observed characteristics, and it is important to evaluate the likelihood that the disclosed similarity supports the hypothesis of the link between cases. This work recommends additional information, including certain technical, contextual and behavioral characteristics that could be collected routinely in cyber-investigations to support similarity computation and link evaluation.

Introduction

Con artists are attracted to the Internet because of the large victim pool, and because of the distance between them and their victims, which reduces the risk of being identified and apprehended. There are an increasing number of online scams, including romance, auction fraud and advanced fee fraud. The ability to find similarities between cases can enable digital investigators to detect some repetition in crime, like in serial offenses committed by the same person or group, and to observe crime patterns or trends that would otherwise be invisible such as online 'hotspots' and repeat victimizations. A crime repetition occurs when crimes are committed by the same offender, target a certain type of victim, employ a common modus operandi, or occur in a particular setting (Cusson, 2012).

Finding similarities between cyber-investigations of online scams can be challenging. Perpetrators frequently change their digital identities and technical tools they use to commit offenses (e.g., email addresses, domain names, URLs, IP address), making it more difficult to find links between related cases. Exact matches of such characteristics may miss important repetitions between cybercrime at both the technical and behavioral levels. Relying on exact matches is also not resilient to inconsistencies in the way information is captured, including data entry errors. There is a need for automated mechanisms to find near similarities in digital traces

* Corresponding author.
E-mail address: timothy.bolle@unil.ch (T. Bollé).

https://doi.org/10.1016/j.diin.2018.01.002

left by offenders' activities (a.k.a. technical characteristics) as well as more complex similarities in context and behavior.

The growing quantity and variety of criminal activities and associated digital traces make it more difficult for digital investigators to discern certain non-exact or non-obvious similarities that can reveal repetitions in cybercrime.

In order to find these patterns, and to avoid linkage blindness (Egger, 1984), there is a need for a centralized case repository with the ability to compute similarities based on traces, context and behavioral information. The present work addresses this need with an automated case linkage process and prototype implementation to facilitate the detection and analysis of these repetitions. This system extends to cybercrime the prior work that demonstrated how non-digital forensic data, including near similarity (i.e. non exact matches) of cases, can be used to detect crime repetition. This process is shown in Fig. 1 and was implemented in the PICAR system (Birrer, 2010; Rossy et al., 2013).

As shown in Fig. 1, the process of developing such a system starts with the acquisition of actual information concerning the crime phenomenon being studied. The integrated information can come from multiple sources and can be of different kinds, including forensic data and situational information, such as spatiotemporal data or a description of the modus operandi (MO). Extending this to the digital realm, Section Recommendations for collecting case information of this paper recommends additional information that could be collected routinely in cyber-investigations. All of the acquired information is then integrated in a structured model ("the memory") that supports various types of analysis, including the detection of relationships between similar cases using near similarity of shoeprints, fingerprints, faces, images and other physical traces, as well as behavioral (MO) and spatiotemporal similarities. The use of forensic data for crime analysis purposes is known as forensic intelligence (Ribaux and Margot, 2003). It is important to differentiate between the investigative context, where the objective is to find information and develop a hypothesis, versus the evaluative context, where the objective is to evaluate the confidence into the hypothesis by testing them against facts and, in the end, be able to present the case in a court of law (Kind, 1994). Applied to the crime intelligence process, the establishment of a link between cases, or entities is a hypothesis. Through the investigation, other information will be used to reevaluate the confidence one could have in the hypothesis. To establish this confidence, it may be necessary to verify the results of some forensic methods.

On the basis of this analysis, decisions could be made at both strategic and operational levels to change the crime environment (Birrer, 2010; Rossy et al., 2013). Observing repetition in cybercrimes can help digital investigators to uncover previously unobserved linkages between a series of related offenses, to study patterns and trends in criminal phenomena, to detect specific vulnerabilities of victims, and to recognize a virtual convergence setting of similar crimes (e.g., increasing use of a new technology or online platform to commit various kinds of crime) (Rossy and Décary-Hétu, 2018). At an operational level, having a group of cases can be more interesting to investigate, in terms of total prejudice, information and resources, in contrary to small cyber cases that may not be worth. In addition, finding nearly similar cases can help digital investigators to solve a new case by directing them to analysis methods that were effective in past cases and can be adapted to the new case (Casey, 2013). The process aims to focus attention and resources on the most prolific offenders and the most problematic offenses.

Structure

This paper begins with a summary of related work, followed by a comparative assessment of different approaches to computing similarity. The important distinction between similarity and the likelihood of a link is discussed. Results of evaluating the prototype system using real world data from 207 cyber-investigations are presented. Due to the different types of cases, the kind and amount of traces captured during the investigation vary greatly. The dataset

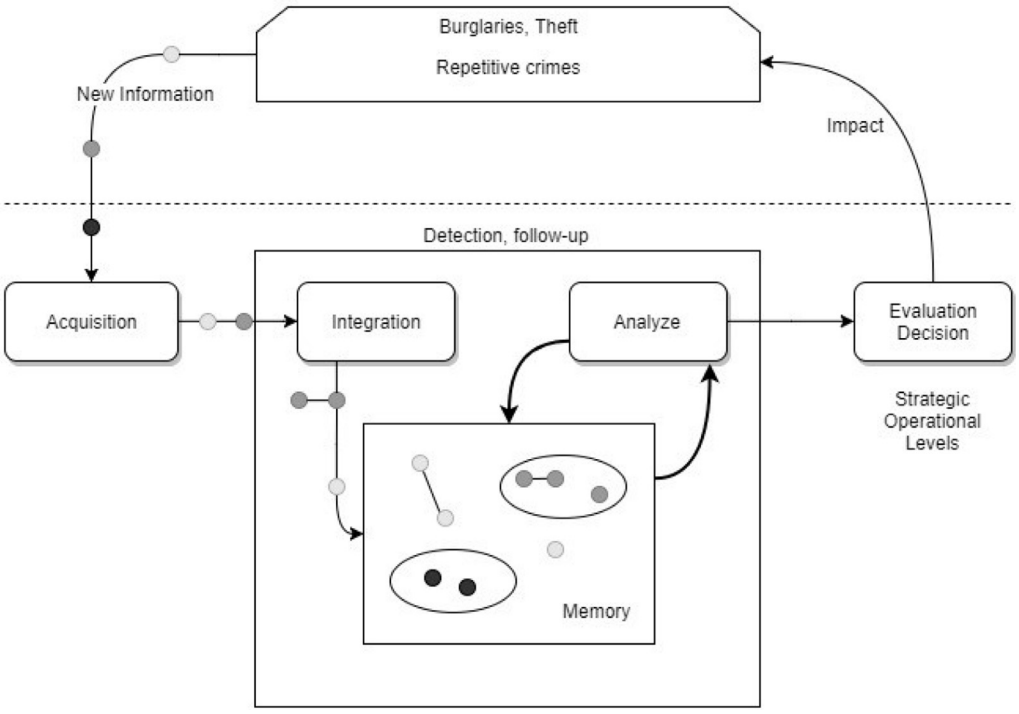

Fig. 1. Systematic crime analysis process (Birrer, 2010).