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

ABSTRACT

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

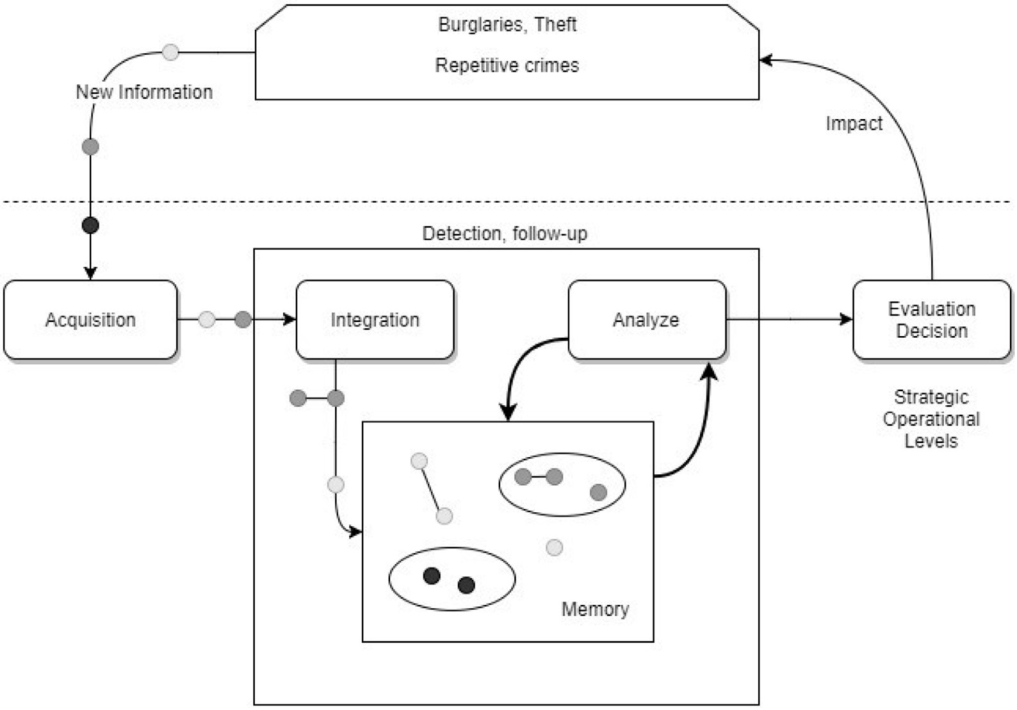

Fig. 1. Systematic crime analysis process (Birrer, 2010).

was analyzed to determine which digital traces were most distinctive and valuable for evaluating linkages and detecting repetitions between cases. This work concludes with recommendations to increase the amount of relevant traces and information captured in future cases to support case linkage. Future work will study repetitions using larger datasets containing a greater variety of digital traces, combining similarity calculations of technical, contextual and behavioral characteristics.

Related research

The process for crime analysis illustrated in Fig. 1 was applied to traditional offense, especially for burglaries, leading to the adoption of the PICAR system to detect repetitions in crimes across different Swiss cantons (Birrer, 2010; Ribaux and Margot, 1999). The PICAR system was created as a collaborative platform that gathers information from multiple Swiss cantons and is accessible by analysts to support case linkage and crime analysis. This system stores links between cases using a combination of information about events, people, vehicles, and near similarities in forensic data such as comparison of shoeprint, fingerprint, faces and DNA. This system supports forensic intelligence processes by storing links amongst traces (i.e. their profiles) but does not store all forensic case data. Especially, the comparison between forensic information takes place in other databases, national or regional. For example, cases can be linked using the fingermark number X from the national AFIS database, without storing precise information about the trace to protect privacy.

NORA (non-obvious relationship awareness) is a complementary approach in criminal intelligence. The limitation of NORA in the context of cybercrime is the prevalence of unstable entities. The fundamentally different challenge in this work is to find potential links between such unstable entities common in cybercrime generally, and cyber fraud specifically.

In order to regroup information from different systems, Albertetti et al (2016) suggest the use of a data warehouse. However, the proposed data warehouse only stores situational information extracted from police reports, and does not introduce the possibility to integrate forensic data.

Concerning internet frauds, Birrer et al (2007), made a preliminary study to show how technical characteristics could be used to link advanced-fee frauds but they only used exact matches. Park et al (2014), analyzed frauds on Craiglist and detected that 48% of the investigated scams were linked to 10 groups. Moreover, they saw that some email addresses were similar (same prefix with different numbers) and that multiple IP addresses came from the same subnet.

In the digital forensic domain, Caltagirone et al (2013) suggest the use of a diamond model to analyze system intrusions in an effort to integrate some technical details in their context with some situational information. This work uses commonality between digital traces and context in related intrusions to support pivoting analysis.

Hutchins et al (2011) also proposed a model to decompose any attack in a succession of actions that allowed the author to perpetrate the offense. This "kill chain" approach can be combined with the diamond model to represent the information available at each step, enabling combined comparison of technical characteristics and action phases.

Systems such as the Malware Information Sharing Project (MISP)[1] and Collaborative Research into Threats (CRITS)[2] have been developed to find potential linkages between attacks against computer systems. These systems use commonalities between indicators of compromise such as IP addresses and domain names, enabling analysts to pivot on a particular atomic piece of information to explore potential linkages. However, these systems do not have automated mechanisms for computing near similarity at the technical, contextual and behavioral levels.

Dietrich et al (2013) propose a methodology using visual hashing of screenshots of malware as near similarity to cluster malware campaigns.

Each of these models have some advantages and weakness, depending of the type of analysis we want to perform. Some are more focused on linking cases, sometimes without storing all the contextual information, while others are more focused on contextual and behavioral analysis of a case.

However, all of these models allow capturing entities or events, and relations between them. In the defense and security context, network analysis, link analysis and, by extension, social network analysis are commonly used (Masys, 2014). As a concrete example, Tyler et al (2003) use emails and the associated email address to detect communities in an organization, and present the possibility to apply their method in a criminal context. These methods support assessment and hypothesis evaluation in many contexts including crime intelligence. As for NORA, these analysis methods work on established networks, and research in these fields usually take into account stable entities. As stated by Tyler et al (2003), one of the major challenges in network analysis is to establish the network and detect redundant entities. This challenge is particularly present in the digital world where, as stated earlier, offenders can easily change their identifying characteristics. Such concealment behavior, as well as inconsistencies in how digital traces are acquired and represented, can result in linkage blindness; failure to detect links between related cases.

Prior work has used graph oriented methods to search for attack patterns in network logs and to detect primary targets or hidden attackers (Wang, 2010). This approach formed graphs on the basis of exact matching technical traces, and did not include near similarity of such traces.

Network analysis and link analysis methods can be applied to the proposed system after repetitions are detected using near-similarity approaches and after the network is established. Future work could also compare the propose method with NORA methods to see the benefits of taking into account unstable entities.

Computing case linkage

At a minimum, any system for finding links between cases should be able to search for exact matches of profiles defined by specific characteristics of the traces. As noted above, such technical characteristics are unstable because they can be changed easily, so there might not be sufficient similarity to link cases with a reasonable level of accuracy. Furthermore, commonality at the technical level does not take into account the multifaceted structure of cybercrime cases, making them weak indicators of overall case similarity. To establish a link and detect repetition across cases with a reasonable level of confidence, it is important to also consider commonality between the trace and the context or behavior. This last aspect can be achieved by the use of network analysis, link analysis and social network analysis which are left for future work.

Near similarity computations

At the most basic level, computing similarity involves matching cases with the most relevant or discriminant trace's characteristics, such as IP addresses, domain names or email addresses.

[1] https://www.misp-project.org/.

[2] https://crits.github.io/.

Computing similarity of technical characteristics is made more complicated by the presence of non-numeric values of varying formats, and partial matches can be necessary. Several measures of similarity are suitable to compare technical characteristics found in digital evidence. In essence, some of these characteristics are multifaceted and their similarity can be calculated using a weighted sum (Casey, 2013). A limitation of these purely technical similarities is that they do not take into account seemingly small differences that are actually very significant for the linkage process. For example, a one-bit difference between two IP addresses could actually correspond to different regions. The IP addresses 73.15.110.251 and 73.16.110.251 are both assigned to Comcast Cable, but the first is allocated in California and the second is allocated in Massachusetts. This type of situation emphasizes the importance of further studying the linkage and its context to evaluate its probative value as discussed in the next section (Linkage Verification, Hypothesis Formation and Likelihood).

Concerning email address, similarities between the usernames should be computed. Indeed, commonalities may suggest the activity of a same offender or group of offenders that will slightly change the addresses they use. This can be true, because their MO require the same kind of addresses or just because is it easier for them to manage the bunch of created addresses. For instance, if a group of authors stole an identity to commit some frauds, they might use different addresses based on this identity for different cases (for instance johndoe@ … and jondoe@ …). Wherein other situations, authors need to send emails that usurp the identity of a well-known organization, like PayPal or AirBnB. In those cases, the email address maybe created using names like “servicepaypal@ …”, “paypalinternational@ …”, ”rentairbnb@ …,” or “airbnbbooks@ …” for instance. Even if the assumption of a same group of offenders could not be proven, the detection of these kinds of MO may be very useful to detect and classify specific kind of situation (i.e. specific online crime phenomena).

These similarities can be computed using usual string similarity algorithms, but a pretreatment is required to clean up the stings by removing some features. For instance, dots, En dashes, Em dashes or numbers should be removed to avoid them to be taken into account during the similarity computation. Among others, Christen (2012) suggest the use of q-gram based methods (Jaccard and cosine similarity), edit distance (Levenshtein algorithm) or algorithms that use both metrics (Jaro and Jaro-Winkler). Depending on the type of similarity we want to highlight, some algorithms may be more interesting. In some systems, the Levenshtein algorithm is employed to help users avoid mistakes by presenting similar email address/domain names.[3]

Algorithm selection

The different algorithms can be evaluated on a test dataset to choose the one that best suits our goals. In a second step, this work evaluates the selected algorithm on real world data. When using real world case data, it was not possible to have a control on the type of traces collected, and on their quality. Consequently, this preliminary study focuses on email addresses.

In order to automatically detect some relevant links, a classifier is required to decide if two address are similar or not. In this study, we decided to use a threshold based classifier. The threshold is chosen to minimize the false negative rate, because in an investigation and crime intelligence processes, we do not want to miss relevant links. False positives can later be excluded during the investigation.

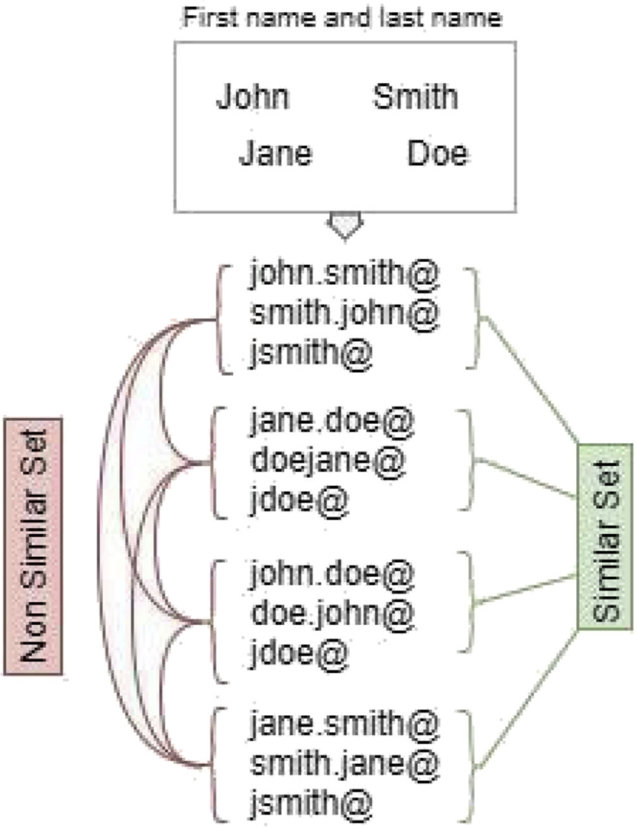

Fig. 2. Test dataset creation process.

The test dataset was composed of a population of non-similar addresses, and a population of similar addresses, both created using random first name and last name (https://www.randomlists.com/random-names). With one first name and one last name, five email addresses were generated with the formats: firstname.lastname@ …, lastname.firstname@ …, flastname@ … (first letter of the first name with the last name), fir.lastname@... (three first letter of the first name and the last name) and firstnamelstnm@... (first name with the consonants of the last name). Two email addresses generated from the same first name and last name were considered similar. The population of non-similar addresses was also composed of addresses were the first name or the last name was common between both addresses. This process is illustrated in Fig. 2. To compute the scores, the Jaccard and cosine similarities were implemented following the algorithms proposed by Christen (2012). A python package was used for the other algorithms.[4]

To compare the different algorithms, we used ROC curves to evaluate their capability to separate the two email addresses populations (Fawcett, 2006). ROC curves show the ratio between false positive rate and true positive rate for different threshold. As we can see in Fig. 3, the Levenshtein metric has the best ratio, meaning that it is possible to find a threshold where we can detect the similar addresses (true positive rate) with few false positives. For this reason, the Levenshtein algorithm was selected for the purposes of this study.

[3] https://github.com/mailcheck/mailcheck/wiki/String-Distance-Algorithms (accessed 11.4.2018).

[4] The code and the test dataset is available here: https://github.com/timbolle-unil/EmailSimilarity.

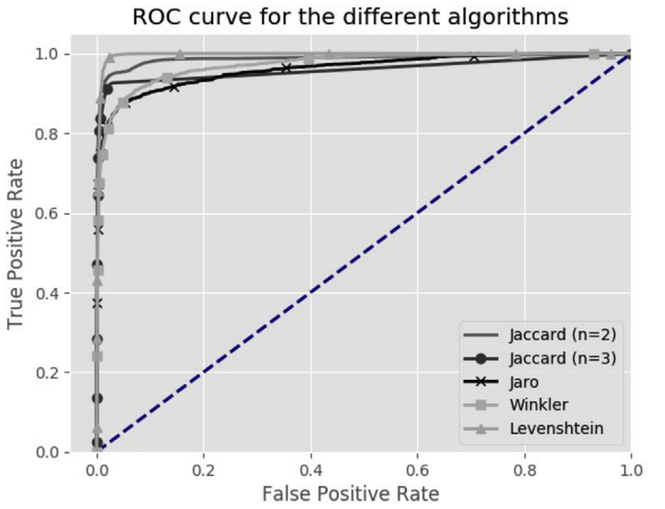

Fig. 3. ROC curve computed for the different algorithms. The cosine similarity for n = 2 and n = 3 gave the same results as the Jaccard similarity with n = 2 and n = 3 respectively.

However, in an investigation and crime intelligence context, the false negative rate should be minimize. Thus, a threshold was fixed to get a false negative rate lower than 1% (Fig. 4).

Basically, this means that we should be able to automatically detect two similar addresses if the score obtained with the Levenshtein algorithm is superior to 0.44. As we discuss in the next section (Linkage Verification, Hypothesis Formation and Likelihood), once the similarity computation is done, an investigator will have to verify if the match is not a false positive. The investigator will then have to evaluate the hypothesis of the link between the two cases.

To compare IP addresses, an analogy can be made with file system and Registry location's similarity evaluation. Similarity of file and Registry locations can be computed, taking into account that it increases from right to left. For example, consider "\Windows\System32\File1.abc" which is more similar to "\Temp\Bin\File1.abc" than "\Windows\System32\File2.exe" even though the latter has more parts in common. Registry locations follow the same pattern. A weighted sum can be used to calculate similarity between file paths of arbitrary length (Casey, 2013). This is an area for future work in cases involving files and Registry values, such as malware investigations. In regards to IP addresses, the similarity increase from left to right. For instance, looking at the subnet part of the address could allow to detect addresses from the same network. From the real world case data obtained for this study, we saw that these traces are not common in cyber frauds, or are rarely collected by the police. The collection of such details and their context could permit more detection of repetitions (see 6. Recommendations for collecting case information).

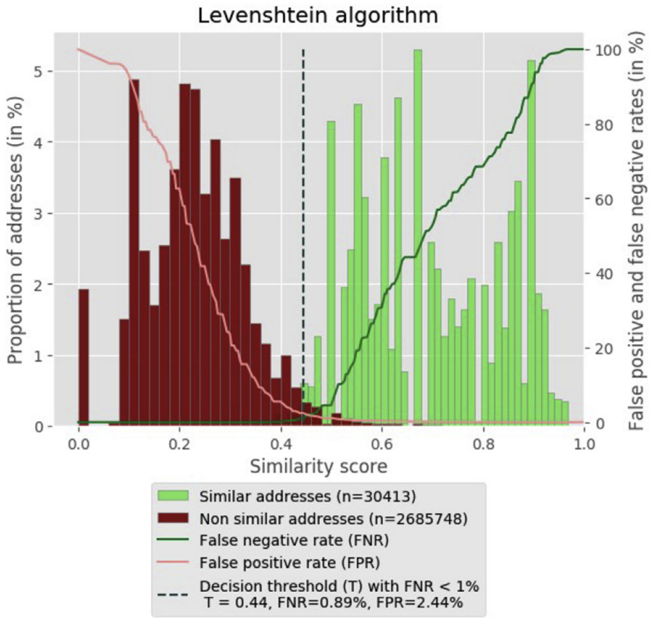

Fig. 4. Distribution of the scores for similar and non-similar addresses with the threshold selection.

Linkage verification, hypothesis formation and likelihood

When designing a system to detect links automatically based on near similarities between digital traces, it is important to treat such links as hypothetical. Such potential links can be useful for making investigative decisions and developing strategies (Kind, 1994). For forensic purposes, it is not sufficient to base conclusions on a weak link, and it is necessary to perform additional analysis to evaluate the likelihood that: 1) a match is a true positive, and 2) the two cases are actually related.

In the first situation, when an automated algorithm computes a high similarity score between two elements, this should be treated as a hypothesis of a link.[5] The digital investigators might then want to tag the detected link as a true positive, confirming that there is a clear similarity between technical elements. To be effective, a system for computing similarities must also allow digital investigators to remove false positive from the system, effectively overriding the automatically computed similarity score. Keeping a log of such user interactions can thus be required to refining the accuracy of the similarity comparison algorithms.

In the second situation, when a digital investigator confirms that a computed link is a true positive, this does not necessarily mean there is a link between two cases. For instance, even with a high score of similarity, if obtained on common words, a digital investigator may not be able to conclude that the same group of offenders is behind two cases because of a lack of confirming information. The digital investigator may conclude that there is a similarity but that it may be explained by other hypothesis (i.e. similar MO, copycats, etc.). This information should however remain in the system, since it might be reevaluated based on the integration of new cases or guide the search for new information.

With this distinction in mind, it is beneficial to give digital investigators the option to input the likelihood of the linkage between two cases. Thus, when taking decision on investigative measures, the investigator will be able to take into account the uncertainty associated to the information of the case. The preferred approach to representing likelihood varies between organizations. Some organizations use a simple likelihood scale (low, medium, high), other groups use scales with five or seven distinctions, and some use likelihood ratios (ENFSI, 2015).

Such kind of evaluation requires context information to apprehend the similarity score. For instance, the IP address found in an email header has a different meaning than an IP address linked in the body of the email. The system should thus integrate the relevant information. This could lead to recommendations for what police should collect in future cases, such as the collection of a complete copy of email messages (with full header details).

[5] Such as the list of possible hits in an AFIS system.

Evaluation

To test the usefulness of those similarities, data from 207 real world cybercrime cases, provided by the digital forensic department of the state police in Geneva, were integrated into the database and processed for link discovery. As shown in Table 1, there were multiple types of offenses in the data, which led to a variety of available information in each case.

The data extracted from the cases were integrated in a MySQL database. We choose a database model that was flexible. The model allows to create links between entities, like pseudonyms and partial identities, email addresses, postal addresses, etc. Contextual information can be added in the link, as well as the confidence level the investigator have in it. From this database, statistics and link charts were computed. The charts were created with the GoJS JavaScript library to visualize links between cases and reconstruct networks. The similarity scores can be computed on the database elements and new links are created if the score is higher than the threshold and validated as true positives by the analyst. An overview of this data management process is available in Fig. 5. Each step of this process can support various technology, depending on what kind of analysis we want.

The information in many of these cases included email addresses but did not include IP addresses or other traces that might be useful for computing case comparison and similarity scores. For this reason, the evaluation concentrated on email addresses.

It was possible to compute 15,400 comparisons between email addresses. Using a threshold at 0.44, 597 address pairs lead to positive results. For the purpose of the evaluation, a manual verification was done to check for false positives and false negatives. It revealed that only 40 of them were truly similar, which means that there was some commonalities between the addresses in the pairs, and that there were no false negatives. The false positive rate here is 3.63%, which is higher than the false positive rate of 2.44% in controlled test data depicted in Fig. 4. This difference might be due to the fact that when the populations were created, not all types of similarity were taken into account and that they did not perfectly represent the real world situation. False positive 3.63% is reasonable in the context of cyber fraud investigations with moderately sized datasets, but might be too high in other contexts involving larger datasets. The threshold can be adjusted to the specific needs of the case and context. A threshold at 0.5 gave only 225 positive results and only one false negative. This demonstrates that changing the threshold may change the number of false positives but as noted previously, the choice of the threshold mainly depends on how many false negatives are acceptable in an investigative context. Future work will concentrate on refining and combining methods to detect repetitions using larger datasets.

This testing indicates that the proposed method is effective for detecting similar characteristics between real world cases. The second step is to evaluate the hypothesis concerning the relation between the cases. This evaluation should include other information collected during the investigation that may change the confidence the investigator has in a hypothesis.

In the following relational diagrams (Figs. 6 and 7), straight lines represent a direct link between two elements, for instance when an email addresses appears in a case. Dashed lines represent links that were detected using the similarity algorithm. In Fig. 6, we can see links between email addresses that were already linked with one or more cases, as for the cases number 78 and 80. This show that the

Table 1
Number of different types of case visible in the data, with the total and average amount lost in Swiss Francs (CHF).

Type of offense	Number of successful cases (Total number of cases)	Average amount lost	Total amount lost
Romance	6 (8)	192′922	1′157′530
Inheritance	1 (4)	37′296	37′296
Lottery	2 (2)	16′972	33′943
Other	5 (11)	8′818	44′091
Buying/Selling of vehicles	7 (10)	7′010	49′071
Identity theft	7 (17)	2′764	19′347
Apartment location	27 (49)	2′308	62′308
Buying/Selling of goods online	56 (83)	1′931	108′151
PaysafeCard (advance fee fraud)	16 (23)	976	15′619

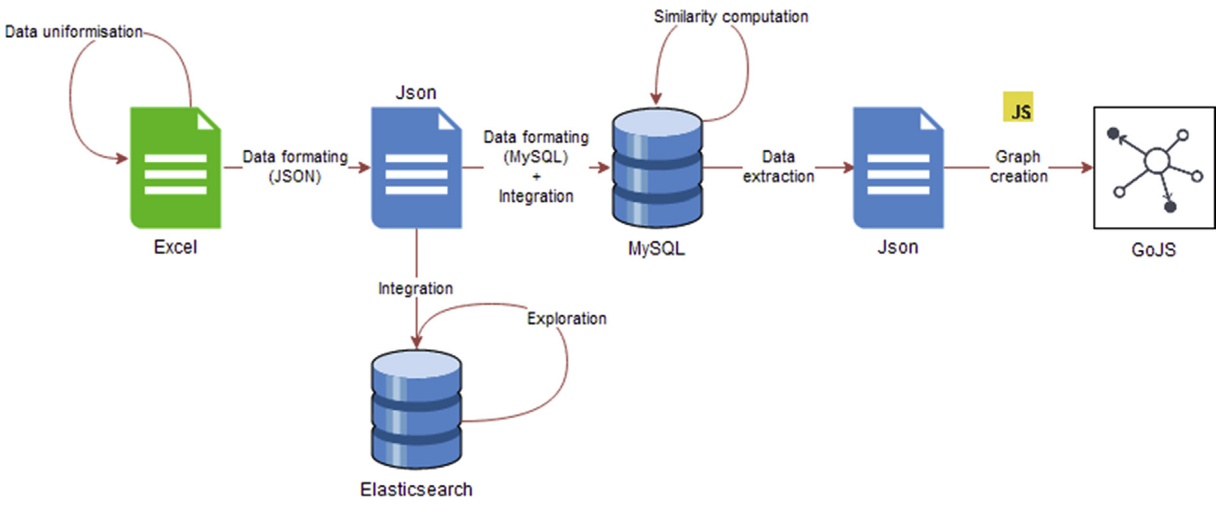

Fig. 5. Data managment process.

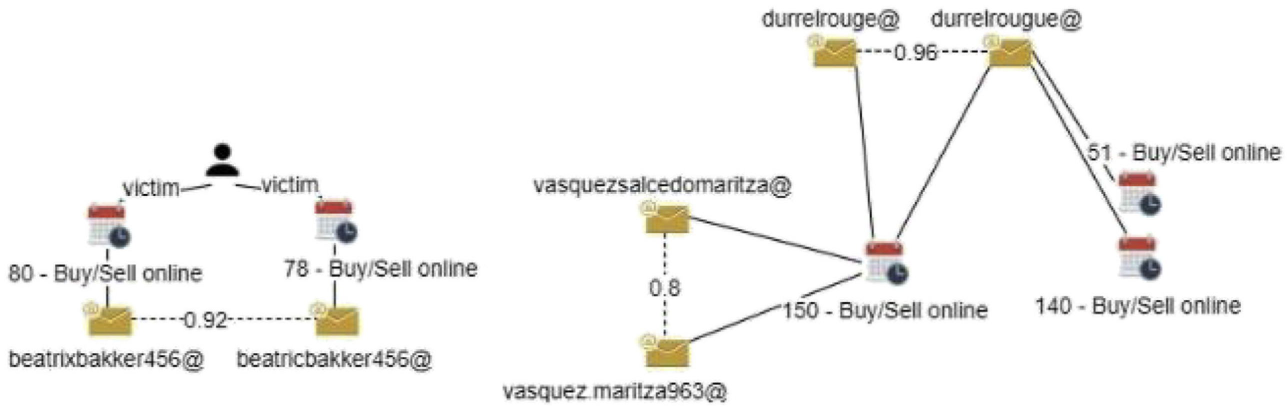

Fig. 6. Relational diagram showing the links between events and other entities, with the link computed with the similarity algorithm. The calendar icon represents a case; the letter icon represents an email address and the person icon represents a person. The straight lines represent known links and the dashed lines represent a link detected with the similarity computation and the numbers on the links are the associated scores.

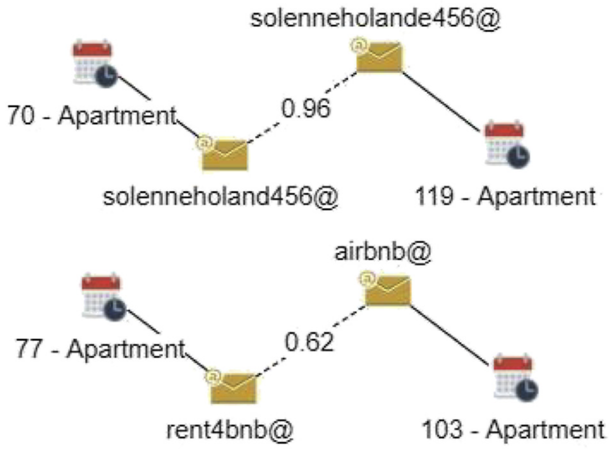

Fig. 7. Relational diagram showing the links between events and other entities, with the link detected with the similarity algorithm.

links created using a similarity algorithm are highlighting already established links. In some cases, as illustrated in the case 150 in Fig. 6, we see multiple email addresses in the same cases that present similarity. This shows that there is concretely some similar addresses that are used in cases, and the similarity are not necessarily input errors.

The near similarity found in the data allow digital investigators to find new links between cases that were not evident before, as shown in Fig. 7.

In both cases, the links based on similarities are useful. In the first case, they confirm the already established links. In the second case, they allow the detection of new links.

In some situation, illustrated with the email addresses found in the cases 70 and 119 in Fig. 7, similarity highlights very close addresses that vary only by a few letters. In this situation, the link might be considered as strong (high confidence in the hypothesis that the two cases are linked) because it seems unlikely to find two email addresses based on a name varying by one letter by chance, when taking into account the geographical area of interest and the popularity of the name in it for instance.

It was also possible to highlight other kinds of similarities where common words are used (e.g., AirBnB in Fig. 7). In those situations, it is not possible to be highly confident in the link between the cases, at least to infer a same offender. Nevertheless, it can lead to detect repetitions at another level (MO, behavioral). In this case, the hypothesis would be different and the detected link should be evaluated regarding this particular hypothesis. The results in Fig. 7 demonstrate the value of computing these near similarities for finding previously undetected relationships, and shows the potential for linking cases on the basis of behavior, not just technical characteristics.

This approach permits digital investigators to find new links between cases and can provide new insights into offender behavior.

Recommendations for collecting case information

The ability to detect links and patterns across cases fundamentally depends on the information that is collected during the cyber-investigation. To increase the consistency and completeness of information available for computing case similarity, police can use formalized guidelines to collect information in cyber-crime investigations:

- Contextual information (e.g., start date and end date, description by victim and witnesses).
- The crime classification, such as online scam, malware attack, identity theft, child abuse material, etc. and a situational classification of the criminal activity (romance scam, investment opportunity, inheritance, etc.). Additional research is needed to define these classifications in regards to the various situational contexts.
- The method used to reach the victim: email, instant messaging, social networking, auction website, chat room, online game, mobile phone application
- Amount lost/paid
- Victim's response to scam (sent money, sent personal details, no response)
- Other witnesses of the offense (person or organization)
- Identity-related information of the offender (e.g., names, addresses, phone numbers, country, birth date, website, or IP addresses) and a description of how the information was obtained.
- Details about bank account (e.g., account number, bank name and location) and other accounts used by the offender to receive funds.
- Type and number of prepaid card used, and any details about how the cards were subsequently used to make purchases (for an advanced fee fraud where prepaid card are used).
- Although it is crucial to collect account details (e.g., email addresses, forum alias), their value can be limited by lack of context. Collecting all correspondences between the victim and

the offender provide valuable context and characteristics. Whenever feasible, full headers of email messages should be collected.

These kinds of information are already partially collected by some institutions but the amount and the type of data collected can greatly vary. For instance, in Switzerland, it is possible to report a crime to MELANI, the Reporting and Analysis Centre for Information Assurance,[6] by sending a free-form textual description of the offense. On the other hand, the Australian Cybercrime Online Reporting Network[7] (ACORN) uses an online reporting form with precise questions to capture more structured information about the offense and its classification, the victim, the suspect, the method and the loss. Ultimately, a combination of structured and unstructured information can be useful for computing linkages and detect repetitions.

Conclusions and future work

This paper presents a new systematic approach to computing near similarly between distinctive digital traces to detected linkages and repetitions across cases in a cyber-investigation context.

The evaluation of the developed prototype indicates that the approach of computing similarity is required to avoid linkage blindness. Future work will explore ways to increase the accuracy of the automatic detection of relevant links in order to enhance its effectiveness in real world cases. The evaluation of the roles of these links to infer crime repetitions should also be done. Crime series reconstruction were presented as example. But the systematic analysis also aims at detecting online 'hotpots', as well as victims' vulnerabilities.

The evaluation in this work concentrated on email addresses to compute similarities between cases. Future work will extend the approach to additional types of digital traces to evaluate their relative value for case linkage. This will require more data and recommendation have been made to encourage and facilitate investigators to collect those data and their context.

There is significant potential for future work in the area of enhancing similarity computation between cyber-investigations. Future work can evaluate effective algorithms for computing similarity in cyber-investigations involving file paths and Registry values such as intrusion and malware investigations. In addition, future work evaluate the effectiveness of combining near similarity approaches (to establish networks) with network analysis methods, including graph analysis, for detecting repetitions in the context of cybercrime investigations.

The methods and technology presented in this work are sufficiently general to be applied in other digital forensic subdomains, including intrusion and malware investigations where finding similar behaviors and indicators of compromise is crucial not only for reusing past digital forensic solutions (e.g., reverse engineering) but also in performing link analysis between related intrusions across multiple organizations. In this context, future work could include comparison of results with the exact matching capabilities within existing system such as Malware Information Sharing Project (MISP) and Collaborative Research into Threats (CRITS).

As we said previously, the enhancement of the similarity accuracy could permit the complete automation of the detection of links across cyber-investigations (still requiring verification and hypothesis testing). The combination of multiple measurement could also enable detection of more specific types of similarity (for instance similarity between names versus similarity between words in email addresses). This may also be achieved by adapting the threshold.

The results of this work also show that finding similarities between some traces' profiles can reveal repeated MO and behaviors. Similarities between cases could also be detected at this higher level, which is interesting to understand particular phenomenon and guide future investigation. Future research can study how similarities can be computed at these higher levels of abstraction.

Acknowledgements

We are grateful to the digital forensic department of the state police in Geneva for motivating this work, giving feedback throughout the project, and providing data for evaluation.

We would also like to express our gratitude to Quentin Rossy and Felix Freiling for their reviews and suggestions of this paper.

References

Albertetti, Fabrizio, Grossrieder, Lionel, Ribaux, Olivier, Stoffel, Kilian, 2016. Change points detection in crime-related time series : an on-line fuzzy approach based on a shape space representation. Appl. Soft Comput. 40, 441–454.

Birrer, Stéphane, 2010. Analyse systématique et permanente de la délinquance sérielle: place des statistiques criminelles; apport des approches situationnelles pour un système de classification ; perspectives en matière de coopération. Université de Lausanne, Lausanne. OCLC : 718150156.

Birrer, S., Ribaux, O., Cartier, J., Rossy, Q., Capt, S., Zufferey, M., 2007. Exploratory study for the detection and analysis of links between prospective advance fee fraud e-mails in an intelligence perspective. IALEA J. 17, 11–21.

Caltagirone, Sergio, Pendergast, Andrew, Betz, Christopher, 2013. The Diamond Model of Intrusion Analysis urlalso. http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA586960 (DTIC Document).

Casey, Eoghan, 2013. Reinforcing the Scientific Method in Digital Investigations Using a Case-based Reasoning (CBR) System. University College Dublin, Dublin.

Christen, Peter, 2012. Data Matching : Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer Science & Business Media.

Cusson, M., 2012. Répétitions criminelles, renseignements et opérations coup-de-poing. Problèmes actuels de science criminelle 21 (2008), 37–52.

Dietrich, Christian J., Rossow, Christian, Pohlmann, Norbert, 2013. Exploiting visual appearance to cluster and detect rogue software. In: Proceedings of the 28th annual ACM symposium on applied computing. ACM, pp. 1776–1783.

Egger, Steven A., 1984. A working definition of serial murder and the reduction of linkage blindness. J. Police Sci. Adm. 12 (3), 348–357.

ENFSI, 2015. ENFSI guideline for evaluative reporting in forensic science (Approved version 3.0) Available at: http://enfsi.eu/sites/default/files/documents/external_publications/m1_guideline.pdf.

Fawcett, Tom, 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27 (8), 861–874.

Hutchins, Eric M., Cloppert, Michael J., Amin, Rohan M., 2011. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. Lead Issues Inf Warf Secur Res 1, 80.

Kind, S.S., 1994. Crime investigation and the criminal trial: a three chapter paradigm of evidence. J. Forensic Sci. Soc. 34 (3), 155–164.

Masys, A.J., 2014. Networks and Network Analysis for Defence and Security. Springer Science & Business Media.

Park, Youngsam, Jones, Jackie, Mccoy, Damon, et al., 2014. Scambaiter: understanding targeted nigerian scams on craigslist. System 1, 2.

Ribaux, O., Margot, P., 1999. Inference structures for crime analysis and intelligence : the example of burglary using forensic science data. Forensic Sci. Int. 100 (3), 193–210.

Ribaux, O., Margot, P., 2003. Case based reasoning in criminal intelligence using forensic case data. Sci. Justice. ISSN: 1355-0306 43 (3), 135–143. https://doi.org/10.1016/S1355-0306(03)71760-2.

Rossy, Q., Décary-Hétu, D., 2018. Internet traces and the analysis of online illicit markets. In: Rossy, Q., Décary-Hétu, D., Delémont, O., Mulone, M. (Eds.), The Routledge International Handbook of forensic intelligence and criminology. Routledge International, Abingdon UK. Routledge International.

Rossy, Quentin, Ioset, Sylvain, Dessimoz, Damien, Ribaux, Olivier, 2013. Integrating forensic information in a crime intelligence database. Forensic Sci. Int. ISSN: 03790738 230, 137–146. https://doi.org/10.1016/j.forsciint.2012.10.010.

Tyler, J.R., Wilkinson, D.M., Huberman, B.A., 2003. Email as spectroscopy: automated discovery of community structure within organizations. In: Communities and Technologies. Springer, Dordrecht, pp. 81–96.

Wang, W., 2010. A Graph Oriented Approach for Network Forensic Analysis. Iowa State University.

[6] https://www.melani.admin.ch/melani/fr/home.html.

[7] https://www.acorn.gov.au/.