Technology assisted analysis of timeline and connections in digital forensic investigations

Hans Henseler*
Magnet Forensics, Waterloo, Canada and University of Applied Sciences Leiden, The Netherlands
hans.henseler@magnetforensics.com

Jessica Hyde
Magnet Forensics, Waterloo, Canada and George Mason University, Fairfax VA, USA
jessica.hyde@magnetforensics.com

ABSTRACT

This article describes ongoing research on the application of AI techniques such as Graph Neural Networks to assist investigators with the discovery of relations and patterns in digital forensic evidence. Digital forensic analysis of smartphones and computers reveals forensic artifacts that are extracted from structured databases maintained by the operating system and applications. Such forensic artifacts are part of a forensic ontology which can be used to build a relational graph of identifiers (e.g. users, documents) and a timeline of events. This information can assist with answering key investigation questions such as who, when, where etc. We propose to use a graph database and query language to assist in this analysis. Further, using key identifiers and aliases we want to augment digital forensic artifacts with entities, relations and events by extraction from the full-text of unstructured electronic contents such as emails and documents.

CCS CONCEPTS

• **Computing methodologies** → **Semantic networks**; *Neural networks*; • **Applied computing** → **Law**; **Investigation techniques**; • **Information systems** → *Users and interactive retrieval*; • **Human-centered computing** → *Visualization toolkits.*

KEYWORDS

Digital Forensics, AI, Link analysis, Timeline, Technology Assisted Discovery, Graph databases, Text Mining, Graph Neural Networks

1 INTRODUCTION

Digital evidence continues to grow exponentially in investigations and prosecution of suspects in both criminal as well as civil cases. Not only in advanced cybercrime investigations, as in, ransomware investigations or as part of incident response, but also through the use of digital forensics in homicide cases or internal (corporate) investigations where the suspect's smartphone and/or laptop needs to be examined. Smartphones and other portable "wearable" electronics leave digital traces that can be linked to persons and locations. The exponential growth of digital traces, as well as the expansion of cybercrime, and digitization of investigative methods represent significant changes to society and lead to a broadening horizon of digital investigation [9].

This article presents work in progress on research that focuses on the use of Artificial Intelligence (AI) as an emerging technology that can assist forensic examiners in the discovery of patterns and relations in digital evidence. It builds further on the ideas presented in earlier work on computer assisted extraction of identities in digital forensics [15], on Semantic Search for E-Discovery [20] and [12], on finding digital evidence in mobile devices [14] and on the link and timeline analysis that is present in modern digital forensic tools [7].

Our vision differs from existing applications of AI in E-Discovery that typically rely on machine learning for classifying digital content such as predictive coding and active learning [11] that filter and cluster emails, chats and documents or classification of pictures with weapons, drugs and nudity. In stead we attempt to apply AI in the discovery of relevant relations in temporal connection graphs that are derived from extracted digital forensic artifacts.

Smartphones and Internet of Things (IoT) devices contain many other digital traces that are a treasure trove in a forensic investigation. Such traces can prove to be more personal than written communication because they do not only reveal our conscious but also our unconscious behavior. Also smartphones have become very personal because of their link with social media and biometric protection (e.g. fingerprint, iris). However, this type of information is machine generated and grows at an even faster pace than our personal communication. Forensic investigations are in need of more effective search strategies that can leverage the richness of detailed forensic in modern digital evidence (e.g. from smartphones, cloud, IoT devices etc.). We propose that investigators are assisted with discovery through using semantic nets that are obtained from digital evidence. We refer to this as *technology assisted discovery* as opposed to technology assisted review that is very common in E-Discovery investigations.

This paper is structured as follows. Section 2 describes digital forensic investigations and the key questions that are relevant when investigating a case. It also describes related work on a digital forensics ontology that can assist when taking a semantic AI approach and explains some use cases why this is helpful when investigating digital evidence. In section 3 we explain modern digital forensic investigations and illustrate how link analysis and time line visualisation are currently assisting forensic examiners in digital forensic investigations. Section 4 presents our vision on how AI techniques such as graph databases and entity extraction can help discovering patterns and relations in these semantic networks of digital artifacts. Finally, in section 5 we present conclusions and identify future research opportunities.

*Corresponding author.
In: Proceedings of the First International Workshop on AI and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2019), held in conjunction with ICAIL 2019. June 17, 2019. Montréal, QC, Canada.

Published at http://ceur-ws.org.

2 DIGITAL FORENSIC INVESTIGATION

Digital forensic investigation typically has three phases: data collection, data examination and data analysis. Data collection involves the correct preservation and copying of digital data sources. Data examination relates to the investigation of copies of digital data sources to find files, extract fragments etc. without interpreting the resultant findings in the context of the case. Data analysis involves the analysis, reconstruction, interpretation and qualification of the evidence which is obtained from the digital data sources. The research proposed here focuses on the analysis of digital evidence.

2.1 Investigation Questions

In any investigation the investigators, regardless if they are senior legal counsel in legal E-Discovery or senior investigating officers or detectives in a criminal investigation, try to answer the following 'golden' investigation questions:

1 Who–was involved?
2 What–happened?
3 Where–did it happen?
4 How–was the crime committed?
5 When–did the crime take place?
6 With what–was the crime committed?
7 Why–was the crime committed?

The analysis of digital evidence in E-Discovery investigations typically focuses on document review and analysis where reviewers and senior investigators analyse textual content. They are assisted by machine learning (also known as predictive coding and continuous active learning [11]) to identify relevant emails and documents to speed up their investigation. Digital forensic investigations on smartphones and computers are a bit different. Here investigators go beyond email and document analysis and study digital artifacts that can be quite pertinent when trying to answer these questions [14].

Who-questions can often be answered by investigating which person is using an e-mail address, user account or phone number. Communication via text messages, chat and email may help to understand what has happened. Call details records, GPS-locations and WiFi-network tell something about the location of a smartphone and consequently of it's user. Pictures and video can provide visual clues how a crime was committed and with what kind of weapon. Date and time of a file or trace, tell when data was last accessed, modified or created. Computers and smartphone maintain detailed records when apps and users were active and which files were involved. Besides messages that a user communicated via emails and chat messages, search history from a browser or specific apps can help understand motive and premeditation.

2.2 Digital Forensics Ontology

Document analysis in E-Discovery heavily relies on the review and analysis of unstructured information that is contained in emails and documents. The analysis of digital forensic artifacts described above is more structured. In order to understand this structure and to be able to analyse it, it is useful to have a digital forensics ontology. The Cyber-investigation Analysis Standard Expression (CASE) [10] provides such an ontology. CASE is an open standard that is currently under development. It can be used to describe different types of digital evidence from various domains such as incident response, counter terrorism, criminal investigations, forensic investigations and gathering of intelligence. CASE enables better coordination of investigations in different jurisdictions so that criminal individuals and organisations are discovered faster while generating a more complete overall view on their criminal activities.

Once a semantic network has been formed based on a digital forensics ontology, it can assist with identifying possible crime scenarios and with testing hypothesis which is becoming increasingly important in investigations. Sometimes it's more important to know with who a victim, suspect or witness communicated, and where these persons were than actually knowing what has been communicated. AI can assist investigators with detecting correlations that can lead to the discovery of relationships that were not known. This is called link analysis. Analysing a social network from a collection of emails is not new but link analysis based on digital forensic artifacts relies on a much richer set of data.

2.3 Use cases

Modern digital forensic tools have a feature that performs *link analysis* to assist forensic examiners with their investigations. Axiom is a commercial digital forensics processing tool created by Magnet Forensics that build a connections database from relationships between discovered artifacts (e.g. users, files etc.). Triples (subject, predicate, object) are extracted following the forensic ontology similar to the CASE ontology introduced in the previous section. These triples define a forensics ontology that is used by Axiom to automatically generate relationship graphs.

Subject	Predicate	Object
file	accessed on	system
file	accessed on	USB
file	accessed by	user id
file	transferred with	program name
file	transferred by	user id
file	related	cloud
file	emailed to	email address
file	downloaded with	program name
file	downloaded by	user id
contact name	contacted with	device
contact name	contacted by	person
picture hit	similar to	picture hit
file/msg	contains	key words
file/msg	references	file name
call log	call to	contact name
user id	used	program name
user id	searched for	key words

Table 1: Subset of triples of the forensic ontology that is used in Axiom

Link analysis has interesting use cases for forensic examiners:

(1) Given a hit the examiner needs to see a visual representation of all related evidence. Where the 'related' links are one of the concepts identified in the forensics ontology.

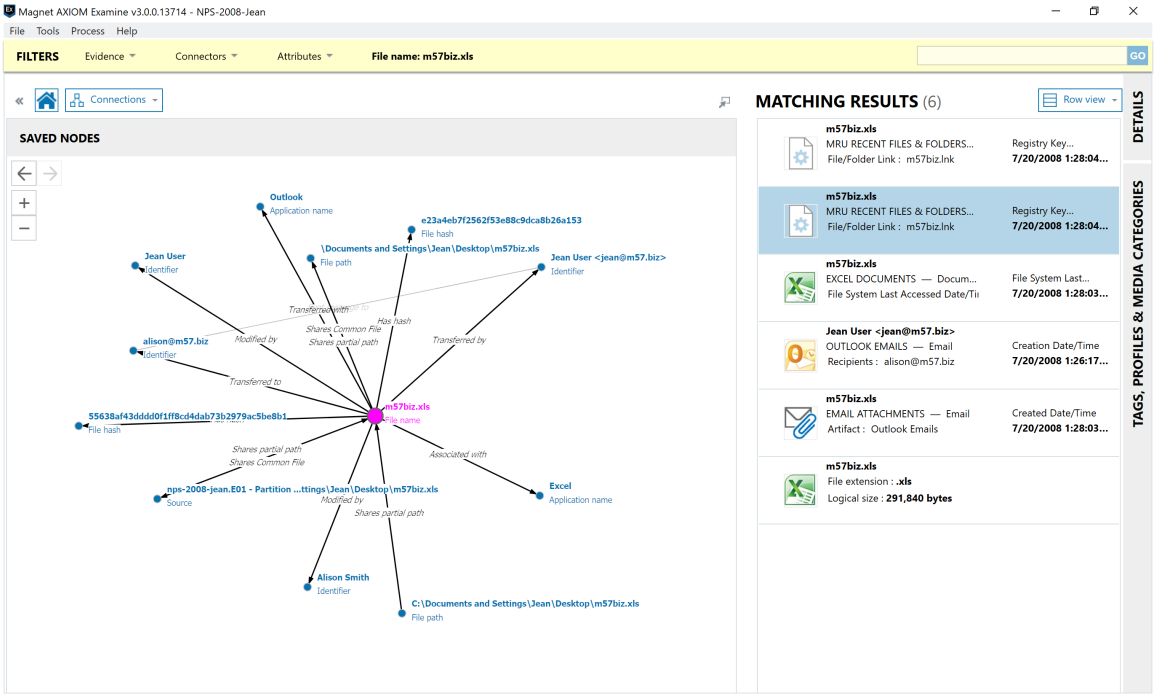

Figure 1: Link analysis example in Axiom on the M57-Jean scenario showing the 'm57bis.xls' file in the center and highlighting one of the records in the Windows MRU (most recently used) list.

(2) Given a link to related evidence the examiner should be able to follow the link and may want to pivot the data around the destination or choose a different visualisation. For example, the examiner identified a search query in browser history and then wants to review all events on the system before and event this query was executed.

Table 1 above lists a simplified digital forensics ontology illustrating a number of triples that make up the key forensics ontology that is used by AXIOM to build a relation graph from digital evidence.

3 EXPERIMENT

To validate the idea and explore the potential of AI for assisting with the discovery of patterns and relations in digital forensics data, we have processed the M57-Jean scenario [6] in Axiom (version 3.0).

The M57-Jean scenario is a single disk image scenario involving the ex-filtration of corporate documents from the laptop of a senior executive. The scenario involves a small start-up company, M57.Biz. A few weeks into inception a confidential spreadsheet that contains the names and salaries of the company's key employees was found posted to the "comments" section of one of the firm's competitors.

The spreadsheet 'm57bis.xls' only existed on one of M57's officers-Jean. Jean says that she has no idea how the data left her laptop and that she must have been hacked. The investigator has been given a disk image of Jean's laptop and is asked to figure out how the data was stolen, or, if Jean isn't as innocent as she claims.

3.1 Axiom Link analysis

Link analysis in itself is not a new concept in digital forensics as is reflected by work published in 2015 [8] and was introduced earlier in 2005 in the field of network forensics [21]. However, it tends to focus on traditional 'call chain analysis'-focusing on phone calls, text messages, and/or social media connections or IP addresses between people or computers rather than the artifacts they create [7].

Artifact relationship analysis goes beyond visualizing relationships between people and computers. It applies the link analysis concept to files and operating system artifacts, helping a forensic examiner to visualize relationships within artifacts and across evidence sources, such as, computers, mobile devices, and even cloud-based accounts.

Figure 1 above presents an example of link analysis in Axiom. This example was discussed in a Magnet Forensics webinar [16]. The tree like structure on the left side shows the file name of a spreadsheet "m57biz.xls". It shows various relations to other elements, e.g., "Transferred by" and identifier "Jean User <jean@m57.biz>", "Hash hash" with a md5 as well as a sha1 hash value, "Application name" relation with an application named "Outlook" etc. The right side of the picture displays matching results. This overview lists records from the Windows MRU (Most Recently Used) list, file system last accessed date, Outlook email record etc. Axiom allows the user to navigate the graph manuall by selecting an end node and making it the center node by double clicking.

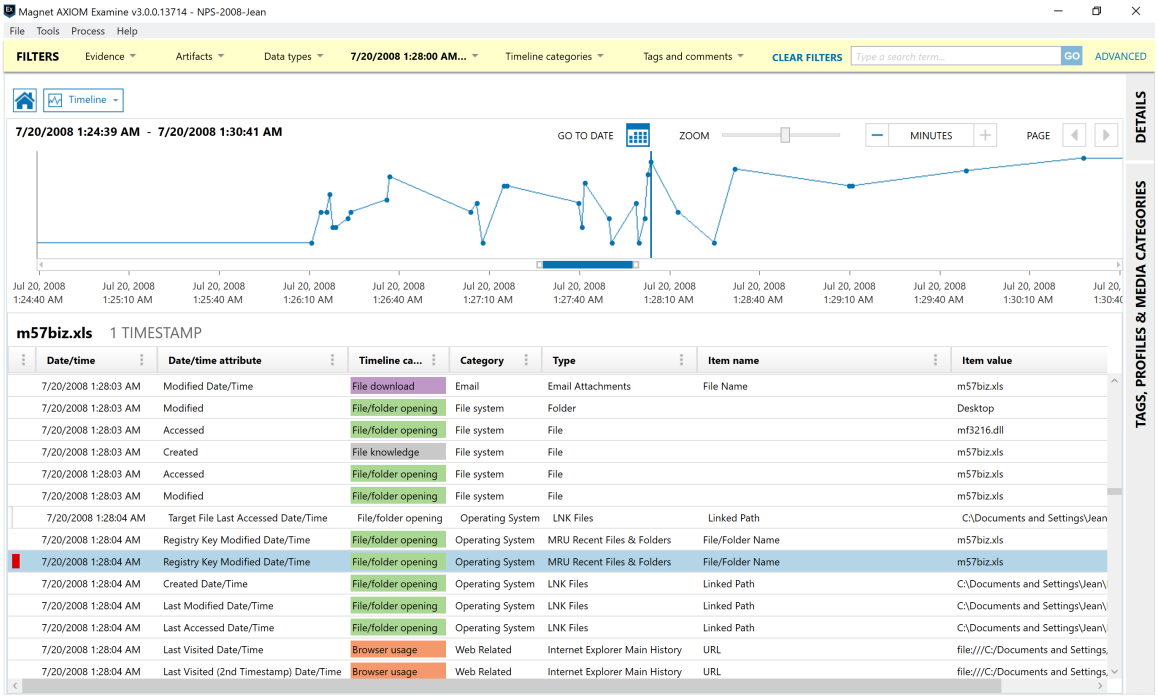

Figure 2: Link analysis illustrating a relative time selection of the MRU artifact highlighted in figure 1

3.2 Axiom Timeline analysis

In [13] an overview is presented of the evolution of timeline analysis in digital forensics. Initially, timeline analysis was focused on file-based dates and times. Around 2010 the first tools became available that started using times from inside files. Modern digital forensic tools (both open source as well as commercial) have advanced timeline capabilities that visualise digital forensic artifacts.

Figure 2 presents a screenshot of the new timeline visualisation and analysis feature in Axiom 3.0. The top section shows a timeline reflecting artifact counts for a period of 6 minutes starting from July 20, 2008 1:24:40am and ending 1:30:40am. The table below the timeline presents a detailed view of the artifacts presented in the graph. At the top is a "File download" record from email, followed by "File/folder opening", then the "File knowledge" reflecting the creation of a new file "m57biz.xls". Such a sequence of artifacts may help understand how a file came into existence on a computer and if it was opened on that computer.

Generation of timelines has also received much attention outside the field of digital forensics. Many applications exist that allow for creation of time lines in an investigation. For example, building case chronologies with CaseFleet [2], create a timeline for your court case with TrialLine [5] and assembling case facts in a chronological order with CaseMap [3]. However, our first impression is that each one of these tools relies on manual development of case timelines without the help of artificial intelligence.

Both timeline analysis as well as link analysis are (separately from each other) considered powerful instruments in an investigation. However, we propose that in combination these features become even more powerful enabling an examiner to analyse links in the relation graph in a chronological order which provides more meaning and context then when simply filtering a timeline for selected entities or filtering the relation graph for a particular time frame.

4 PROPOSED RESEARCH

Our research focuses on the combination of timeline and link analysis. In order to accomplish this we propose to use a graph database with a graph query language. The graph can initially be constructed from forensic artifacts. With modern graph databases and graph query languages it becomes easy to augment this graph with additional data. This could include data from non-digital sources but also by text mining the full-text of electronic documents and emails, new relations might be uncovered that previously would have required human inspection of the contents of such documents.

4.1 Graph database and language

Visualisation of traces in a network, on a map or on a timeline can assist a forensic investigator to understand the story that is behind the data. By ingesting the information that is extracted by Axiom in the M57-Jean case in a graph database, it becomes possible to experiment further with visualisations and discovering relations.

Cypher is a graph query language that allows for expressive and efficient querying of graph data [1]. It lets developers write graph queries by describing patterns in the data. If we have a graph describing our digital forensic artifacts, Cypher is designed to be a human readable query language and is suitable for both developers as well as forensic examiners.

Cypher describes nodes, relationships and properties as ASCII art directly in the language, making queries easy to read and recognize as part of your graph data. Figure 3 below presents an example of a simple Cypher query.

Cypher is supported by a variety of graph databases. We intend to use Neo4j [4] for our experiments which will start with modeling a relational graph based on a selection of the digital forensics ontology that is used by Axiom. Then we'll investigate how easy it is for examiners to formulate Cypher queries and which standard queries can be formulated to identify interesting relationships that can be prioritized for review.

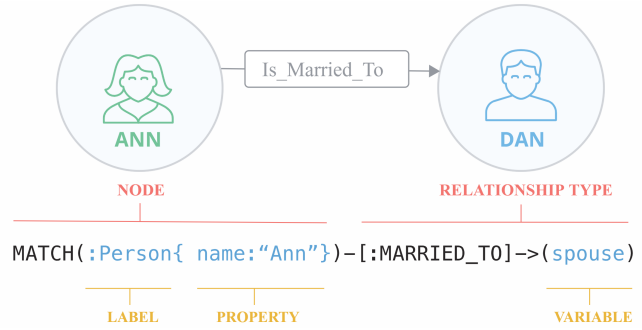

Figure 3: Example of a Cypher query

4.2 Integration with other information sources

Once the digital artifacts from a case have been imported in the graph database, it becomes quite simple to add relations and objects in the same case that were discovered through other sources. These can either be other sources of digital evidence, e.g., other cases, call detail records, or from non-digital sources such as witness and victim statements, lawful interception, observation, open source intelligence or case time lines that were manually created assisted by software such as mentioned in paragraph 3.

By leveraging the scalability of modern graph databases a great variety of additional information can be included in the automated analysis [18]. Further research is required to investigate what other information (that is typically available in a criminal investigation) can be combined with the digital forensics graph in a useful way. We expect that to some extent even scenarios and hypothesis can be formulated as (a set of) graph queries which can be tested against the graph containing all known information on the case.

4.3 Extracting relations and timelines from full text

More than 90% of the information around us is mostly unstructured, e.g., documents, emails and chat messages. Text mining can help investigators by turning this unstructured information into structured data. Entity extraction can extract entities (e.g. names of people, organisations, places etc) and events from full text. Unfortunately the extraction of entities is error prone and generates many false positives making the results useless. By using identifiers that have been discovered from digital forensic analysis the entity extraction can be targeted reducing the number of false positive identities.

Some interesting work in the field of entity-centric timeline extraction has been reported in [17]. A prototype tool is being developed that can extract structured information on events for a given entity of interest and place anchors on a time line for these events. It uses massive streams of textual documents as input (e.g. online news, social media posts or any crawled web documents).

With digital forensics it is already possible to extract identities from structured information through digital forensic analysis [15]. When an examiner identifies an interesting identity probably this identity will have associated email aliases, accounts, phone numbers etc. Once this information is known it can be added to the relation graph and will help in extracting a timeline of events that are related to these identities.

4.4 Using AI to understand graphs and timelines

Analysing a graph using a visualisation tool seems simple enough. As graphs get bigger, traditional mathematics can help with the analysis of the graph but these methods also have their limitations. One of the problems is that there is no clear beginning or ending of a graph (assuming it's cyclic) and that large scale matrix operations that are typically required for graph analysis do not compute due to memory and time restrictions.

Ontologies, graph database and graph query language are well established and are hardly considered AI techniques. Extracting relations and timelines from full text are well established AI techniques that we hope to leverage in our research but we have no intention of improving this. The core idea in our innovation is to use Graph Neural Networks (GNNs) as a new AI technique that can assist with the analysis of large time-based graphs of relations.

GNNs were first introduced in 2009 [19] and have recently gained increasing popularity in various domains, including including social science (social networks), natural science and knowledge graphs [22]. Similar to the successful application of Convolutional Neural Networks (CNNs) in image classification and Recurrent Neural Networks (RNNs) in natural language processing, variations of GNNs have have demonstrated ground-breaking performance on many tasks.

Our research hypothesis is that we can use GNNs to model interesting relation graphs which can assist investigators with the identification of relevant subgraphs from a highly complex case graph that is automatically constructed from digital forensic artifacts combined with other case data.

5 CONCLUSIONS

We propose to use a graph database and query language to assist in digital forensic investigations. We start with a relation graph that is based on connections from digital forensic artifacts. Further research and experiments are needed to study how forensic examiners can interact with this graph and how to extend the graph with other data sources. In particular we intend to study how events on a timeline can be added to the graph, how information from non-digital evidence can be added and how we can improve the performance of existing entity extraction techniques on unstructured

data from emails and documents. Finally, we want to research if new machine learning techniques such as GNNs can be used to learn from investigators what link and event patterns are interesting from an investigator perspective.

REFERENCES

[1] [n. d.]. About Cypher. Adapted from https://www.opencypher.org/about, Accessed: 2019-04-22.

[2] [n. d.]. CaseFleet: Building Powerful Case Chronologies with CaseFleet. Company website, https://www.casefleet.com/timelines-case-timeline-software, Accessed: 2019-05-22.

[3] [n. d.]. CaseMap: Chronology best Practices. Product pagee, https://www.casesoft.com/download/chrons.pdf, Accessed: 2019-05-22.

[4] [n. d.]. Neo4j: The Internet-Scale Graph platform. Neo4J website, https://neo4j.com/product/, Accessed: 2019-04-23.

[5] [n. d.]. TrialLine: Legal Timelines for Your Court Case. Company blog, https://blog.trialline.net/legal-timelines-for-your-case-in-court, Accessed: 2019-05-22.

[6] 2012. M57-Jean Scenario. In *Digital Corpora.* Scenario published at the Digital Corpora website, https://digitalcorpora.org/corpora/scenarios/m57-jean, Accessed: 2019-04-15.

[7] 2018. Telling the Story of Digital Evidence. (2018). Magnet Forensics blog, https://www.magnetforensics.com/blog/telling-the-story-of-digital-evidence/, Accessed: 2019-04-22.

[8] Fergal Brennan, Martins Udris, and Pavel Gladyshev. 2015. An Automated Link Analysis Solution Applied to Digital Forensic Investigations. https://doi.org/10.1007/978-3-319-14289-0_13

[9] E. Casey. 2017. The broadening horizons of digital investigation. *Editorial of Digital Investigation* 21 (2017), 1–2.

[10] E. Casey, S. Barnum, R. Griffith, J. Snyder, H. Van Beek, and A. Nelson. 2017. Advancing coordinated cyber-investigations and tool interoperability using a community developed specification language. *Digital Investigation* 22 (2017), 14–15.

[11] Gordon V. Cormack and Maura R. Grossman. 2015. Multi-Faceted Recall of Continuous Active Learning for Technology-Assisted Review. In *SIGIR 2015 Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[12] David Graus. 2017. *Entities of Interest.* Ph.D. Dissertation. Informatics Institute, University of Amsterdam.

[13] C. Hargreaves and J. Patterson. 2012. An automated timeline reconstruction approach for digital forensic investigations. *Digital Investigation* 9 (2012), S69–S879.

[14] H. Henseler and V. Noort. 2017. Finding Digital Evidence in Mobile Devices. (2017). Presentation at DFRWS US 2017 conference, https://www.dfrws.org/conferences/dfrws-usa-2017/sessions/finding-digital-evidence-mobile-devices, Accessed: 2019-04-03.

[15] Jop Hofste, Hans Henseler, and Maurice van Keulen. 2013. Computer assisted extraction, merging and correlation of identities with Tracks Inspector. In *Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL 2013).* 247–248. https://doi.org/10.1145/2514601.2514639 Demo-paper.

[16] J. Hyde. 2018. Connecting the Dots Between Artifacts and User Activity. (2018). Recorded webinar, https://www.magnetforensics.com/resources/connecting-artifacts-property-theft-webinar/, Accessed: 2019-04-23.

[17] Jakub Piskorski, Vanni Zavarella, and Martin Atkinson. 2018. On the Development of an Entity-Centric Timeline Extraction Tool. 821–824. https://doi.org/10.1109/ASONAM.2018.8508798

[18] G. Sadowski and P. Rathle. 2016. Why Modern Fraud Detection Needs Graph Database Technology. (2016). Neo4j blog, https://neo4j.com/blog/fraud-detection-graph-database-technology/, Accessed: 2019-04-22.

[19] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The Graph Neural Network Model. *Trans. Neur. Netw.* 20, 1 (Jan. 2009), 61–80. https://doi.org/10.1109/TNN.2008.2005605

[20] David van Dijk, David Graus, Zhaochun Ren, Hans Henseler, and Maarten de Rijke. 2015. Who is involved? Semantic search for e-discovery. In *ICAIL 2015 Workshop on Using Machine Learning and Other Advanced Techniques to Address Legal Problems in E-Discovery and Information Governance (DESI VI Workshop).*

[21] W. Wang and T. Daniels. 2005. Network Forensics Analysis with Evidence Graphs. (2005). Published in the proceedings of the DFRWS US 2005 conference, https://www.dfrws.org/sites/default/files/session-files/paper-network_forensics_analysis_with_evidence_graphs.pdf, Accessed: 2019-04-22.

[22] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018. (12 2018). Research gate, https://www.researchgate.net/publication/329841448_Graph_Neural_Networks_A_Review_of_Methods_and_Applications, accessed: 2019-05-31.