Applying Semantic Technologies to Fight Online Banking Fraud

Rodrigo Carvalho*†, Michael Goldsmith*, Sadie Creese*
*Cyber Security Centre, Department of Computer Science
University of Oxford, Oxford, UK
{rodrigo.carvalho, michael.goldsmith, sadie.creese}@cs.ox.ac.uk
†Brazilian Federal Police, Brasília, Brazil
carvalho.rac@dpf.gov.br

***Abstract*—Cybercrime tackling is a major challenge for Law Enforcement Agencies (LEAs). Traditional digital forensics and investigation procedures are not coping with the sheer amount of data to analyse, which is stored in multiple devices seized from distinct, possibly-related cases. Moreover, inefficient information representation and exchange hampers evidence recovery and relationship discovery. Aiming at a better balance between human reasoning skills and computer processing capabilities, this paper discusses how semantic technologies could make cybercrime investigation more efficient. It takes the example of online banking fraud to propose an ontology aimed at mapping criminal organisations and identifying malware developers. Although still on early stage of development, it reviews concepts to extend from well-established ontologies and proposes novel abstractions that could enhance relationship discovery. Finally, it suggests inference rules based on empirical knowledge which could better address the needs of the human analyst.**

***Keywords*-ontology; malware; cybercrime investigation; digital evidence;**

I. Introduction

Tackling cybercrime is a growing challenge for Law Enforcement Agencies all over the world. According to the Interpol site, "More and more criminals are exploiting the speed, convenience and anonymity of the Internet to commit a diverse range of criminal activities that know no borders, either physical or virtual." [1]

Due to its inconsistent detection occurrence, it is very difficult to measure global cybercrime activity such as phishing scam precisely. Even so, some organisations monitor the evolution of attacks reported by specific sources across the year. Among other interesting findings in its first-quarter 2014 report [2], the Anti-Phishing Working Group (APWG), an international coalition that brings together several relevant institutions affected by cybercrime, states that:

- The number of phishing sites leaped by 10.7 percent over the fourth quarter of 2013;
- The number of phishing attacks observed in Q1 was 125,215. That is the second highest number of sites detected in a first quarter, eclipsed only by the 164,032 seen in the first quarter of 2012;
- Payment Services continued being the most targeted industry sector.

Probably one of the reasons for such alarming statistics is the complexity of investigation. Pieces of evidence from a single offence might be spread in servers and computers across different countries. To make matters worse, an United Nations report on cybercrime affirms that "widespread reliance on slow-moving traditional mechanisms such as mutual legal assistance, the emergence of country cooperation clusters, and a lack of clarity on permissible direct law enforcement access to extraterritorial data present challenges to an effective global response" [3]. Even if all countries agreed at once in sharing cybercrime information, there would still be a vocabulary barrier to overcome: different technical and legal terms used by each one would make data integration a non trivial task.

In addition, due to its "novel" investigation-skills requirements, there are not enough trained law enforcement officers to tackle cybercrime accordingly, which certainly increases the work backlog. Moreover, the capable ones might get overloaded with the amount of data to be analysed, hampering the discovery of relationships and patterns.

Therefore it is imperative to use the available computer processing power in a more intelligent way, in order to better exploit all the cybercrime evidence stored in either open or closed sources. An important first step would be the "construction of a common language and a set of basic concepts about which the security community can develop a shared understanding...a common language and agreed-upon experimental protocols will facilitate the testing of hypotheses and validation of concepts." [4]

A. Online Banking Fraud Investigation

Bank customers are among the most common targets of phishing scam attacks. This is particularly true in Brazil, where the fast development of the online banking sector was not accompanied by adequate public-awareness campaigns regarding its risks and necessary precautions. As a consequence, a vast Online Banking Malware (OBM) cybercrime ecosystem has emerged.

In it, the same malware is normally sold to and used by multiple thieves, who outnumber developers by a significant factor. Therefore, arresting the latter could have a major cascade effect in reducing fraudulent transactions. As a secondary, but not less important consequence, more effective

978-1-4799-8657-6/15 $31.00 © 2015 IEEE
DOI 10.1109/EISIC.2015.42

OBM investigations could dissuade developers from malware programming.

The fact that many criminal organisations use malware from a single developer might increase the chances of finding relevant leads. Thus, it is imperative to correlate and reason about such horizontally sparse evidence found in multiple seized devices.

B. Ontologies

One possibility towards a more intelligent use of computer resources is to provide them with semantic capabilities. This can be achieved by creating a knowledge base, in which data is stored together with its human-attributed meaning.

Initially, an ontology should be developed. According to [5], an ontology is "a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application."

By establishing a common vocabulary about a domain's concepts and relationships which is understandable by both human and computer agents, an ontology enables, among other features [6]:

- To share a common understanding of the structure of the information among people and software agents: once agreed and implemented, cybercrime data from distinct sources (e.g. law enforcement agencies) would become compatible, even if not shared yet; additionally, this could make training human agents and developing software systems more homogeneous and integrated;
- To enable reuse of domain knowledge: there are many common concepts relevant to distinct domains (e.g. file hashes are important for both online banking fraud and chain of custody). If such concepts are well implemented and maintained, they can be extended to different ontologies;
- To make domain assumptions explicit: it is easier to notice, understand and change domain assumptions when they are integrated with the data and defined using common language constructs. On the other hand, knowledge maintenance gets more complex if "raw" data is separated from its meaning, statically embedded in the software source code.

In addition to automated computer reasoning, the adoption of an ontology would enable data input reliability, easier information sharing and homogeneous training and software development between different actors.

C. Objective

This paper will propose the initial version of an ontology whose main objectives are:

1) To map different criminal organisations and identify the malware developers, by uncovering relationships between a great quantity of supposedly unrelated evidence;
2) To facilitate future cybercrime data integration and improve related discussion among non-technical people, by providing a standardized way of collecting, storing and representing information;
3) To aid evidence discovery, by enabling forensic analysts to consult the knowledge base for leads on current unknown evidence, such as the names given to text files from a specific malware variety.

In addition, this paper provides some inference rules that could help cybercrime investigation. It also discusses forensic concepts to be extended from existing ontologies and proposes novel ones.

II. Related Work

In recent years, many researchers acknowledged the benefits of ontologies, which may have caused their migration from the realm of Artificial Intelligence laboratories to the desktops of domain experts [6]. A notable example is the Semantic Web [7], an attempt to better integrate data from disparate sources on the Internet so it can be shared and reused more rationally. Additionally, fields like biology and medicine are also exploring its potential [8].

Different tools and methodologies have been proposed to support such initiatives. The following sections will cite some of them, as well as review related papers.

A. Articles

One of the first attempts to employ semantic technologies in the criminal domain was proposed in 2005, in the paper "Ontology-based decision support system for crime investigation processes" [9]. According to its authors, such framework would optimize information collection, storage, processing and exchange, in order to better support decisions regarding "the knowledge of the crime scene investigation tactics and strategies of various types of crimes and their peculiarities, where to look for traces, what investigation plan to make up and what problems to solve.". Although providing a systematic description of crime investigation workflows and suggesting an ontology representing general crime concepts, it does not include cybercrime and the related digital evidence in its scope.

Then, in 2009, a Cyber Forensics ontology was proposed [10], linking the different subclasses of the concepts "Law", "Crime Case", "Criminal", "Crime Type" and "Evidence", the latter further describing collection procedures. By presenting a top-level approach regarding any crime that could leave digital evidence, it does not delve into the analysis of the evidence content itself, but focuses on the medium in which it was found (e.g. a memory stick or a hard drive). Notwithstanding the importance of digital evidence categorization, our Online Banking Malware Ontology (OBMO) aims at relating and reasoning upon the content of the digital evidence in addition to its metadata.

In 2013, researchers from the Computer Emergency Response Team (CERT) at Carnegie Mellon University published a paper discussing an ontology dedicated to malware analysis

based on established vocabularies and taxonomies [11]. Their goal is to provide a more scientific approach to malware research, and they hope other experts will adopt such ontology, thus starting to "speak the same language".

Although sharing some goals, the OBMO focuses on the cybercrime investigation rather than the analysis of malware itself, (e.g. it considers the entities responsible for the malware development and use against online banking fraud victims). In addition, it uses a different ontology language, which will be explained in Subsection II-C.

The ontology proposed in this paper is unique in the way it merges some of these topics with the investigation needs and evidence analysis performed by a Law Enforcement Agencies, enabling a task-driven ontology-developing process.

B. Methodologies

The implementation of the complete life cycle of an ontology development process is not a simple task. It involves many concepts inherent to software engineering projects, such as resources management, evaluation and testing, activities scheduling and iterative cycles.

Therefore, analysing the different ontology-engineering methodologies (such as Methontology) in order to define the best fit for the OBM investigation domain is not one of the objectives of this paper. For related information, please refer to the NeOn project report [12] that, besides suggesting a new methodology for building ontology networks, also presents a good evaluation between well-established ones.

This paper will follow the Knowledge-Engineering Methodology steps described by University of Stanford researchers [6], as it favours explaining ontology-specific concepts and issues to the detriment of describing a complete and formal engineering process, thus enabling a better understanding of such technology.

C. Tools

Two of the most complete and advanced ontology-building languages are the Web Ontology Language (OWL) [13] and the Nepomuk Representation Language (NRL) [14]. Although they have common origin (both extend the Resource Description Framework (RDF) [15] and the associated RDF Schema (RDFS) languages) and purpose (both represent and process knowledge in a machine-interpretable way), they initially targeted different domains.

OWL was designed to provide semantic capabilities to the Web, allowing automatic processing and integration of data from distinct sources based on its meaning. It became a World Wide Web Consortium (W3C) recommendation in 2004. In contrast, the objective of the NRL was to provide semantic power to desktop applications, by structuring the context of all personal information stored on someone's computer.

Another important distinction between them is that reasoning in OWL is based on the Open World Assumption (OWA), in which the absence of a statement doesn't mean it is false. Instead, its truth value remains undefined, as there might exist unknown information that could directly affect the assertion. On the opposite, NRL is based on the Closed World Assumption (CWA), meaning that any statement which does not hold a true value is considered false. It matches the "...expectations of the (NRL) users better, as a local desktop is indeed a closed world with a limited, known, processable number of files" [16].

The very nature of cybercrime investigation suggests that it is more adequate to reason upon OWA (after all, the current lack of incriminating evidence in one device does not necessarily means a suspect is innocent, as future forensic analysis might confirm he is guilty). However, the capabilities that semantic technologies could bring to computer forensics and cybercrime investigation will be explored using NRL, as there is a well established group of ontologies from which the OBMO could extend many concepts. Further details will be discussed in the next Section.

Finally, the Protege ontology editor was chosen to implement the concepts, properties and relationships of the OBMO. It is a popular free software that counts on both active community support and tailored plug-ins which make the creation process easier.

III. OBM Ontology

The OBMO proposed in the following subsections follows the methodology described in [6]. Domain-specific information was based on both the first author's computer-forensic experience within the Brazilian Federal Police and also on the content of reports from other forensic analysts.

The domain of the ontology is the digital data stored on devices seized during operations against online banking fraud. Therefore, as a first assessment about which concepts to include on the ontology, 30 forensic reports from devices analysed across 2013 were assessed in terms of:

- Authorship diversity: the selected reports must be from distinct authors;
- Content richness: it may vary significantly from one device analysis report to another;
- Case variety: the selected reports must cover both ATM fraud complaints from different banks and computer devices seized during distinct operations.

After selecting a group of ten reports that met these three criteria, the different types of evidence found in each of them were compared in a spreadsheet. Table I lists some of the prevailing evidence:

Bank ATMs	**Criminal suspects' computers**
Configuration files with keywords	Source code with keywords
USB devices connection logs	C&C servers URLs/IPs
Telephone numbers/SMS	Mail spam lists
Fake ATM screen pictures	IM logins and chat history
Malware filenames and paths	
Email messages	
Victims account details	
People names and nicknames	

TABLE I: Common evidence found in official forensic reports.

In addition, the inquiries from the investigation teams regarding the suspect devices were also considered. They

assisted in the definition of competency questions, which help determining the requirements of the ontology. Some examples are:

- Are there any connections between criminal organisations using the same malware?
- Do these different malware come from the same developer?
- Which are all the email addresses used by a specific criminal organisation?
- Is the malware able to communicate? If so, how and to whom?
- Is there any indication that this suspect is a member of a specific criminal organisation?

A. *OSCAF ontologies*

The Open Semantic Collaboration Architecture Foundation (OSCAF) ontologies [14] provide high-level knowledge representation for digital data concepts such as files, contacts and messages (through the Nepomuk File, Contact and Message Ontologies - NFO, NCO and NMO respectively). It is an excellent foundation for the OBMO for two reasons:

1) Clear alignment of concepts: after all, cybercrime evidence is mostly found within digital data;
2) It counts on long term support and a rich development history.

Furthermore, any schema for cybercrime investigation must necessarily represent the agent committing the offence. To achieve that, an important concept from the Friend of a Friend (FOAF) ontology [17] will be extended: it defines an agent as a "thing" (a person, group or software) that does things. This is particularly useful for the proposed ontology, as it considers malware an entity that does things (e.g. creating text files and sending emails). Further details will be discussed in Subsection III-B.

The following subsections will discuss the application of some OSCAF concepts into the digital forensics and investigation domains. In addition, we propose novel concepts that could better integrate both domains, and suggest related semantic queries that could optimize evidence finding.

B. *Entity*

Inspired on the "agent concept" from FOAF (a thing that does things), the OBMO implements the class *Entity*, containing the concepts Person, Group and Malware, as depicted in Figure 1.

The actual malware is considered an entity due to its capability of taking actions based on the feedback from the environment, which resembles the ones performed by a "real" thief: presenting a bait (phishing scam) to deceive a naive person, writing down the victims' bank details (appending them to text files) and delivering a list containing multiple victims' information to the gang chief (sending emails or Short Message Service (SMS) messages.)

One might argue that the existence of multiple copies of the same malware within the OBM crime ecosystem would undermine their bonding with specific organisations. However, they can be individualized by a combination of fuzzy hashes [18] (either the main files and the auxiliary or embedded ones) and the message recipient in its Portable Executable (PE) file, and by that, be considered a member of an organisation.

Therefore, inference rules such as: "If two malware have significantly different fuzzy hash values, they belong to different families. If the same malware are sending information to the same recipient, then they are members of the same organisation" might apply. In addition, we would know that this organisation is using more than one malware family. If both hypotheses are applied to a big knowledge base, patterns and relationships may emerge. More details will be discussed in Subsection III-D5.

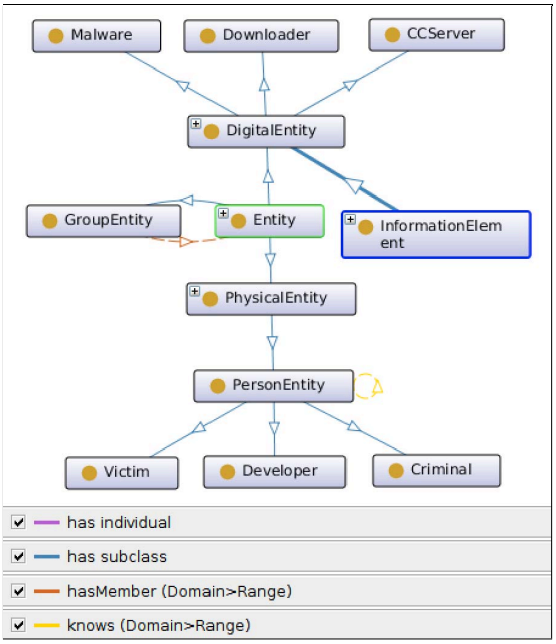

Fig. 1: Entity class and subclasses.

It is still necessary to distinguish between physical and digital entities, as there are intrinsically different ways to describe and relate them: while someone might have a *postal address* and *know* another *PersonEntity* (the symmetric and reflexive relationship denoted in yellow in Figure 1), a piece of software can be identified by its *hash*, and might contain clues about its developer in the PE file (such as the project compilation folder from unfinished malware). This is the reason why *Digital Entity* is also considered a subclass of *InformationElement*, further detailed in Subsection III-D5.

C. *ContactMedium*

In addition to phone calls, which are excluded from the scope of this paper, OBM group members contact each other through email, SMS and IM messages. Therefore, the *ContactMedium* class was extended from NCO, as it implements the *PhoneNumber*, *PostalAddress*, *EmailAddress* and *IMAccount* subclasses.

Moreover, as the OBMO considers malware as an *Entity*, the *TCPIPAddress* class was created. Its main purpose is to represent communication between *DigitalEntities*. For instance, the inference rule from III-B could be expanded to: "If malware from the same organization are downloaded from distinct servers, then the latter are also members from such organisation". Ultimately, this information could help identifying organisations sharing resources with each other.

Figure 2 shows the *hasContactMedium* object property between *Entity* and *ContactMedium*. Also, that any *Message* can be *from*, have a *recipient* or *reply* to an *Entity*'s *PhoneNumber* (in the case of SMS), *EmailAddress*, *PostalAddress* (restricted to *PhysicalEntities*), *TCPIPAddress* (restricted to *DigitalEntities*) or *IMAccount*. The latter has a crucial role in establishing the *know* object property among *PhyisicalEntity* instances, and will be discussed in Subsection III-D4.

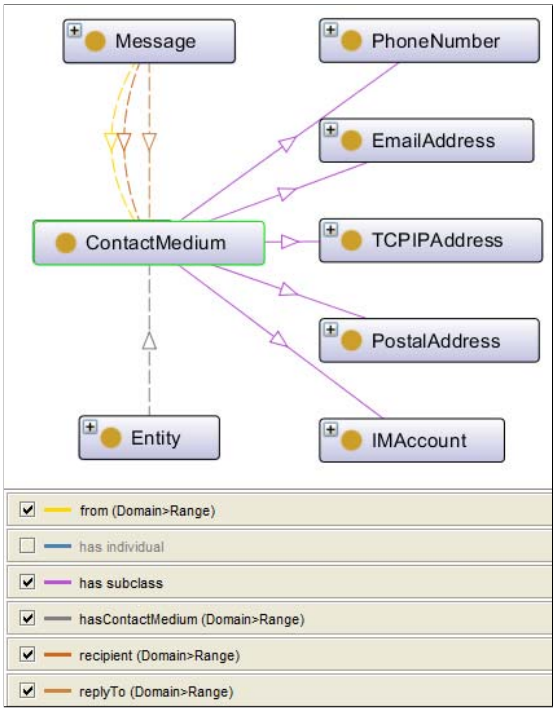

Fig. 2: ContactMedium class and subclasses.

D. Information Element

The *Message* class mentioned in Subsection III-C is a child from the *InformationElement* class, along with its siblings *Contact, ContactList, DigitalEntity and Document*. The OSCAF group of ontologies makes a clear and appropriate distinction between an *InformationElement* (describing content-specific information) and a *DataObject* (representing the "physical" container). They are connected through the *isPartOf* relationship: a *File is*(the physical)*PartOf* a *TextDocument*, for instance.

This approach provides the necessary level of flexibility for describing digital evidence found in seized storage devices, and is one of the main reasons why OSCAF was chosen as the foundation ontology for the OBMO. The NIE specification [14] gives the example of the mailbox, an *InformationElement* subclass that can be represented by either a local *FileDataObject* (e.g. "inbox.pst") or a *RemoteHostAddress* (in the case of the IMAP protocol). Although having different representations, the interpretation (mailbox) is the same.

1) Message: Represents communication within the OBM ecosystem:

- Regular email and IM between criminals: in addition to sender and recipient usernames, identifiable information in the messages content such as names, nicknames and locations could be added as *contact* attributes;
- Regular email and SMS between malware and criminals, containing victims' bank details in text format;
- Phishing emails between spam senders and victims, which contains a link to an evil *URL*;
- Control signals between different types of malware: e.g. a remote C&C server signalling that a victim has just connected to the online banking site, or a downloader requesting a malware from a remote server.

2) Document: A great part of evidence documents encountered during OBM devices forensic analysis are *represented* as text files: instructions for malware usage, email addresses listings, banking information (name, account number, passwords), source codes and others.

Often they include specific strings (e.g. comments, slang, specific characters sequences) that, if *interpreted* using the *keyword* data property, could help relating a great number of documents towards its *producer* identification. For instance, the following inference rule could apply: "If two documents share 4 unique keywords, then they have the same contributor."

Similar properties, such as *definesClass*, *definesFunction* and *definesProgrammingLanguage* could be also related against different *sourceCodes* to help confirming a supposed unique origin.

3) Contact: The *Contact* class extended from OSCAF ontologies is broader than a simple person representation within a IM software, for instance. It encompasses every piece of information that can help identifying an entity. Some examples are *nicknames* found in text files and malware *versionNumbers*. This approach benefits OBM investigation as it allows collecting and reasoning upon little, atomic information dispersed over different cases.

For instance, an unknown *nickname* found in a suspicious message would be inserted into the knowledge base as an instance of the *Contact* class. Because that file's *DataObject* is linked to the owner (*PersonContact*) of the seized device (*DataSource*), that would automatically suggest a weak, yet possible, relationship with the unknown person referred by that *nickname*. This hypothesis could be later confirmed or refuted with further added information.

Finally, there are distinct data properties for *PersonContact* and *MalwareContact*. The former contains *fullName* and *birthDate*, and the latter *versionNumber* and *targetBank*, for instance. However, their *DataObject* representations might

share metadata properties such as *modifiedDate*. Further details will be discussed in Subsection III-E.

4) ContactList: The main reason for extending the *ContactList* is to map which *PhysicalEntities* from the OBM ecosystem know each other. For instance, the *creator* of a *ContactList* already has a weak relationship with the *owner* of the *DataSource* (the seized device), if they refer to different *PersonContacts*.

In addition, the relationship *containsContact* lists all *ContactListDataObjects* found in that device belonging to the specified list. Each one of them would be interpreted as either a new *ContactMedium* instance, or associated to an existing *PersonContact*. As further *ContactList* information collected from different sources is added to the knowledge base, the chances of finding the *PersonContact* associated to a recurrent *ContactMedium* increases.

5) DigitalEntity: Despite being able to do things, a DigitalEntity is still a software. Therefore, it is also a subclass of *InformationElement*, inheriting object properties (e.g. *isStoredAs*, linking it with the corresponding *FileDataObject*), and data properties such as *contributor*. The task of classifying malware based on its *keywords* could be enhanced by dedicated tools like Yara [19], which identifies and classifies malware families based on textual and binary patterns found in the PE files.

It is worth noticing that any *emailAddress* or *phoneNumber* found in its executable would not be *keywords* from an *InformationElement*, but the actual *ContactMedium* from the current *DigitalEntity*. Nevertheless, they could still be compared to find relationships between different malware.

The *Digital Entity* dual nature could make the task of linking related organisations and identifying malware developers easier. After all, automatically merging information from its relationship among other *Entities* with evidence from its *Information Element* class (regarding both its PE file and the metadata of the operational system it was found in) could reduce the amount of suspects to consider.

Figure 3 illustrates an inference flow example. The main distinction between malware from different organisations is the *ContactMedium* to which it sends data to. If such information is unknown, the related *Downloader URL* might also indicate some level of relationship. In addition, the output from previously mentioned fuzzy hashing and malware classification techniques could support inference decisions regarding identifying malware developers.

E. DataObject

The *DataObject* is the container of an *InformationElement*. As explained in [14], "It represents a native structure the user works with. The usage of the term 'native' is important. It means that a DataObject can be directly mapped to a data structure maintained by a native application. This may be a file, a set of files or a part of a file." The relevant subclasses to the OBMO are:

1) CarvedDataItem: Stores information retrieved from the file system non-allocated space. It is created by the forensic recovery tool, and would solely indicate that the content container has been permanently deleted. Although its offset could be easily determined, it does not carry enough investigation relevance to be represented in the ontology.

2) FileDataObject: Comprises files from allocated disk space, whether local, deleted (to the trash bin), remote or embedded ones. It's the most common container for *InformationElements*, and contains relevant linking-capable properties such as *hashValue*, *dateModified* and *fileSize*.

3) RemotePortAddress: As stated in the mailbox example, it is a *DataObject* interpreted as the IP/Port and timestamp of a specific access to a malware downloader server. This information, extracted from *PhishingScam* emails and thieves devices, can help to map infrastructure shared by different criminal organisations along time.

4) ContacListDataObject: Stores each *Contact* within a specific *ContactList*. It needs a specific representation because there might be multiple *ContacListDataObjects* stored in the same *FileDataObject* (e.g. "contacts.edb"). It contains relevant investigation metadata such as the date each contact was added.

F. DataSource

The *DataSource* represents the "root physical container" (e.g. laptop, smartphone) from which information was collected. In addition to correlating devices and cases, this class is also relevant for managing the chain of custody.

After all, it is imperative to assure the device integrity along its way to the court. Because different people manipulate it, starting at the seizure location, passing by the agency's storage room and finally reaching the forensics lab, reasoning upon object properties such as *MovedBy* and *NextDestination* could help identifying suspicious behaviour.

G. Facets and Instances

Facets are not classes, but restrictions applied to both object and data properties' values. According to [6], the most common ones are value type, cardinality and classes domains and ranges. Table II lists some OBM properties along with their facets. Column "C" states either single or multiple cardinality, and the "*" symbol distinguishes data properties rows.

Domain	Property	Range / Type	C
Info.Element	keyword *	string	M
Info.Element	contributor	Contact	M
Info.Element	creator	Contact	1
Entity	hasContact	Contact	M
Contact	hasName	Name	M
Name	nickname *	string	M
Info.Element	isStoredAs	DataObject	M
DataObject	lastModified *	dateTime	1
DataObject	dataSource	DataSource	1
ContactList	containsContact	ContactListObj	M
Info.Element	relatedTo	DataObject	M
Entity	hasContactMedium	EmailAddress	M
EmailAddress	emailAddress *	string	1
Message	from	ContactMedium	1
Message	inReplyTo	Message	M

TABLE II: Some OBMO classes, properties and their facets.

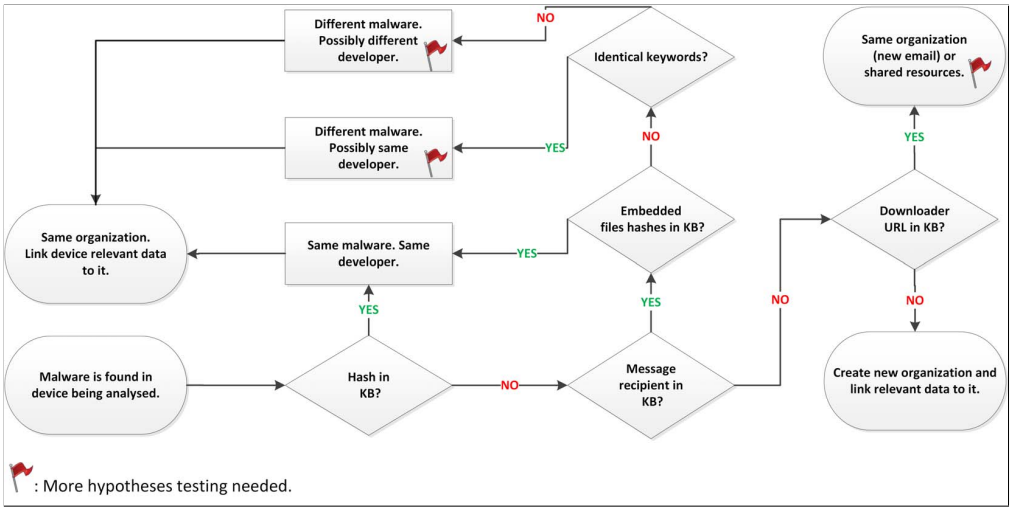

Fig. 3: Inference flow example.

IV. Discussion and Future Work

Semantic technologies could have a positive effect in the way cybercrime evidence is currently analysed, as they merge basic reasoning capabilities with computer processing power. An automated process of relating superficial information from multiple sources would allow the human analyst to analyse and reason upon deeper, more complex information.

Although based on empirical knowledge and official information from current analysis reports, the concepts, relationships and semantic queries proposed in this paper have not yet been tested against real data.

In addition, the classification of the malware as an entity has not been discussed among forensics analysts and investigators. A working knowledge base could help assessing its potential and, if deemed valid, the ontology itself could help disseminating this novel idea.

Thus, the next steps of this work in progress are:

- To implement a working prototype with a reduced set of concepts following a well-established ontology-engineering process;
- To load real data to the knowledge base in order to validate the effectiveness of the classes and inference rules in relationship finding and hypothesis testing;
- To evaluate which pattern matching tools would better suit the ontology needs for correlating files with multiple keywords.

Finally, the feasibility of automatic information extraction has to be considered, as it is a previous important step towards evidence gathering. For instance, in the case that a big chat history file is found, containing long conversations with multiple recipients, how would information (e.g. location, nicknames, email addresses and references to victims' bank details) be collected?

Whereas this problem might be considered out of scope, a failure in addressing it would risk the ontology adoption, as the amount of necessary work to manually input all this data could discourage some users. Nevertheless, an alternative approach for the cases that NLP techniques are not effective should also be considered. Therefore, some future work suggestions are:

- To research Natural Language Processing (NLP) techniques to automate entity extraction and provide content-based file categorization;
- To develop a forensic analysis assisting application that would allow the user to input and tag evidence to the knowledge base effortlessly.

V. Conclusion

This paper has discussed some current issues related to cybercrime investigation which affect law enforcement agencies, mostly derived from the complexity in finding and relating OBM evidence within the great amount of data sent to forensic analysis.

An Online Banking Malware ontology is proposed: its classes, properties and relationships have been thoroughly discussed, and sample semantic queries, based on common tasks performed by forensics analysts and investigators, have been suggested.

This work-in-progress paper expects to spark stakeholders' interest in semantic technologies, as it considers they will be very relevant in future cybercrime tackling.

Acknowledgment

The first author's DPhil programme is funded by CAPES-CSF and supported by the Brazilian Federal Police.

References

[1] "Cybercrime - INTERPOL." [Online]. Available: http://www.interpol.int/Crime-areas/Cybercrime/Cybercrime

[2] APWG, "Phishing activity trends report - 1 st quarter 2014," Tech. Rep., 2014. [Online]. Available: http://docs.apwg.org/reports/apwg_trends_report_q1_2014.pdf

[3] Steven Malby and Robyn Mace, "Comprehensive study on cybercrime," United Nations Office on Drugs and Crime, Tech. Rep., 2013. [Online]. Available: http://www.unodc.org/documents/organized-crime/UNODC_CCPCJ_EG.4_2013/CYBERCRIME_STUDY_210213.pdf
[4] JASON Program Office, "Science of cyber-security," The MITRE Corporation, Tech. Rep., 2010.
[5] Tom Gruber, "Ontology (computer science) - definition in encyclopedia of database systems." [Online]. Available: http://tomgruber.org/writing/ontology-definition-2007.htm
[6] N. F. Noy, D. L. McGuinness, and others, "Ontology development 101: A guide to creating your first ontology," 2001. [Online]. Available: http://protege.stanford.edu/publications/ontology_development/ontology101.pdf
[7] "Semantic web - w3c." [Online]. Available: http://www.w3.org/standards/semanticweb/
[8] "SNOMED CT." [Online]. Available: http://www.ihtsdo.org/snomed-ct/
[9] D. Dzemydiene and E. Kazemikaitiene, "Ontology-based decision support system for crime investigation processes," in *Information Systems Development*, O. Vasilecas, W. Wojtkowski, J. Zupani, A. Caplinskas, W. Wojtkowski, and S. Wrycza, Eds. Springer US, 2005, pp. 427–438. [Online]. Available: http://dx.doi.org/10.1007/0-387-28809-0_37
[10] H. Park, S. Cho, and H.-C. Kwon, "Cyber forensics ontology for cyber criminal investigation," in *Forensics in Telecommunications, Information and Multimedia*. Springer, 2009, pp. 160–165. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-02312-5_18
[11] D. A. Mundie and D. M. Mcintire, "An Ontology for Malware Analysis." IEEE, Sep. 2013, pp. 556–558. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6657289
[12] M. C. Surez-Figueroa, "D5. 4.1. NeOn methodology for building contextualized ontology networks," 2014. [Online]. Available: http://www.neon-project.org/deliverables/WP5/NeOn_2008_D5.4.1.pdf
[13] "OWL web ontology language overview." [Online]. Available: http://www.w3.org/TR/owl-features/
[14] "OSCAF ontologies." [Online]. Available: http://www.semanticdesktop.org/ontologies/
[15] "RDF schema 1.1." [Online]. Available: http://www.w3.org/TR/2014/REC-rdf-schema-20140225/
[16] M. Vlkel, *Personal knowledge models with semantic technologies*. BoDBooks on Demand, 2010. [Online]. Available: http://digbib.ubka.uni-karlsruhe.de/volltexte/documents/1453712
[17] "FOAF Vocabulary Specification." [Online]. Available: http://xmlns.com/foaf/spec/
[18] J. Kornblum, "Identifying almost identical files using context triggered piecewise hashing," *Digital Investigation*, vol. 3, pp. 91–97, Sep. 2006. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1742287606000764
[19] "YARA - The pattern matching swiss knife for malware researchers." [Online]. Available: http://plusvic.github.io/yara/