future internet

Article

Improving Forensic Triage Efficiency through Cyber Threat Intelligence

Nikolaos Serketzis [1,*], Vasilios Katos [2], Christos Ilioudis [3], Dimitrios Baltatzis [4] and Georgios Pangalos [1]

[1] Department of Electrical & Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[2] Department of Computing and Informatics, Bournemouth University, Poole BH12 5BB, UK
[3] Department of Information and Electronic Engineering, International Hellenic University, 57400 Thessaloniki, Greece
[4] School of Science & Technology, International Hellenic University, 57001 Thermi, Greece
* Correspondence: nserketzis@auth.gr

Received: 5 June 2019; Accepted: 15 July 2019; Published: 23 July 2019

Abstract: The complication of information technology and the proliferation of heterogeneous security devices that produce increased volumes of data coupled with the ever-changing threat landscape challenges have an adverse impact on the efficiency of information security controls and digital forensics, as well as incident response approaches. Cyber Threat Intelligence (CTI)and forensic preparedness are the two parts of the so-called managed security services that defendants can employ to repel, mitigate or investigate security incidents. Despite their success, there is no known effort that has combined these two approaches to enhance Digital Forensic Readiness (DFR) and thus decrease the time and cost of incident response and investigation. This paper builds upon and extends a DFR model that utilises actionable CTI to improve the maturity levels of DFR. The effectiveness and applicability of this model are evaluated through a series of experiments that employ malware-related network data simulating real-world attack scenarios. To this extent, the model manages to identify the root causes of information security incidents with high accuracy (90.73%), precision (96.17%) and recall (93.61%), while managing to decrease significantly the volume of data digital forensic investigators need to examine. The contribution of this paper is twofold. First, it indicates that CTI can be employed by digital forensics processes. Second, it demonstrates and evaluates an efficient mechanism that enhances operational DFR.

Keywords: digital forensics; digital forensic readiness; threat intelligence; threat hunting; forensic triage

1. Introduction

Cyber Threat Intelligence (CTI) focuses on the collection and analysis of information concerning current and potential attacks that threaten the security of an organization or its assets. It is a relatively new approach for securing information systems that aims to improve the inefficiencies of traditional defence mechanisms by contributing to the provision of managed security services. As such, CTI is an essential component that offers both proactive and reactive security (also known as incident response and forensics) to protect systems against a plethora of attacks, even against initially undetected compromises, while shortening the window between compromise and detection if protection mechanisms fail [1].

Despite the robustness of CTI security implementations, the proliferation and sophistication of new threats, such as malware and advanced persistent threats (APTs), challenges their effectiveness. For example, Symantec found that "zero-day" exploits are currently in circulation, most commonly on the dark web, for 300 days on average before identification [2], while the mean time of malware instances being undetected on compromised systems is 101 days [3]. Mandiant's report [3] also indicates that four out of ten organisations do not detect such compromises themselves, which can increase incident discovery and response complexity.

At the same time, the new era of data protection legislation, such as in the case of the European Union where the stricter law framework of the General Data Protection Directive (GDPR) has already been enforced, incident response and investigation time become even more critical as undetected compromises and the subsequent incident response may have an adverse financial effect on organisations. Moreover, the complexity, heterogeneity, correlation and sheer volume of data are some of the challenges digital forensics now face [4]. To this extent, the emergence of Digital Forensic Readiness (DFR) is considered a promising approach towards addressing several of these challenges.

Driven by the need to further improve the effectiveness and applicability of DFR, this paper demonstrates that coupling DFR practices with CTI could successfully improve Key Performance Indicators (KPIs) related to the volume, correlation and complexity problems that traditional digital forensics approaches face. The contribution and focus of this paper are illustrated in Figure 1. Considering an investigation scope and all digital evidence (S), the evidence related to a given case would be subset E, whereas subset T represents the evidence collected during the triage phase. In an ideal scenario, the triage set will contain information only related to the case, filtering out all the "noise". However, in a real case scenario, following the execution of the triage process, the following outcomes are expected:

- A subset of digital evidence items identified by triage (true positives or "hits")
- A subset of items not relevant to the case, but included in the triage subset (false positives)
- A subset of items relevant to the case, but not identified by the triage (false negatives or "misses").

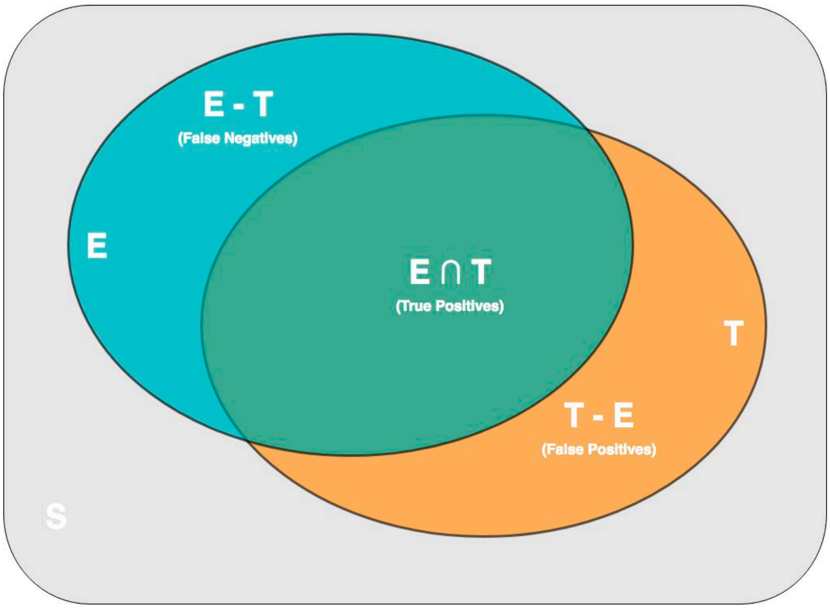


Figure 1. Problem definition.

Due to the nature and purpose of the triage, there is a bias and preference towards reducing false negatives ($\min|\{E–T\}|$) at the expense of increasing the false positives. In other words, a forensic analyst would be in favour of influencing the triage criteria in order to include as much digital evidence as possible rather than risk missing them. In this paper, we argue that this trade-off can be improved by bringing correlation forward to the triage phase.

Building on our proposed threat intelligence informed DFR model [5,6], this paper evaluates its effectiveness by empirically evaluating it through scenarios based on real attack data. The results of the experiments are indicative of the contribution of an enhanced DFR model, and they indicate that it is capable of identifying and limiting the key causes of security incidents in an efficient period of time.

The rest of this paper is structured as follows: Section 2 summarizes previous research about DFR. Section 3 provides an overview of the research conducted by the authors and the model they proposed. Section 4 describes the evaluation methodology, including the setup of the model in a real-world setting. Section 5 presents, evaluates and discusses the results of the experiments. Finally, Section 6 summarizes the benefits of this research and explores future research directions.

2. Related Work

2.1. Digital Forensic Readiness

Almost two decades have elapsed since Tan [7] introduced the term digital forensic readiness to describe the need for discovering new methods to anticipate incident response thought, maximising the usefulness of incident evidence data, while minimising the cost of forensics during incident response. The emergence of that term highlighted the overlap between information security and digital forensics and signalled new ways of thinking about the former, meaning that digital forensics significantly influence the way security is planned, implemented and measured in an organisation [8].

The benefits of DFR have officially been recognized with the adoption of the ISO27043 standard [9] that formalised the way digital forensic investigations proceed across incidents that involve digital evidence. The readiness phase of this standard includes processes like identification of possible digital evidence sources, pre-incident collection, storage, manipulation and analysis of digital evidence and incident detection.

2.2. Digital Forensic Readiness Models

For a long time, enhancing DFR has been one of the main concerns among researchers. For example, Endicott-Popovsky et al. [10] stated that information systems must be designed following an updated 4R (Resistance, Recognition, Recovery and Redress) approach of the NIST Information Systems Development Life Cycle [11]. Grobler et al. discerned that achieving DFR requires the coordination of multiple factors, including people, processes, technology, policies, governance and law [12]. Elyas et al. endorsed such a multifactor approach and also highlighted the relationships between those factors. Their work also stressed the need for proactive data collection preservation and triangulation, indicating that pre-incident technical procedures are essential for effective post-examination [13].

2.3. Digital Forensic Readiness Operations

Concentrating on operational approaches for implementing DFR, Al-Mahrouqi et al. [14] proposed a network forensic readiness and security awareness framework that aims to preserve, analyse and extract relevant knowledge. However, they did not provide any implications concerning its efficiency. The consideration that many organisations increasingly tend to use flexible Bring Your Own Device (BYOD) policies, which in turn can significantly affect their overall information security and DFR posture, guided Kebande et al. [15] to work towards proposing a DFR framework that incorporates honeypot technology to collect, preserve and store Potential Digital Evidence (PDE). Despite its innovativeness, that paper spent little effort proactively analysing PDE and so requires improvement. Similar approaches that employ benign botnets with the aim of achieving DFR on cloud services were proposed by Kebande et al. [16–18]. The idea behind that concept included the installation of obfuscated software on clients' cloud virtual appliances that managed to collect and transmit PDE to pre-configured databases that operated under the Infrastructure as a Service (IaaS) architecture. The prototype they built indicated the applicability of their research, which was primarily focused on proactively gathering PDE, which, in turn, is considered a required step towards ensuring DFR.

Many researchers agree that fostering proactive forensics requires proper elaboration of the obtained PDE [19–23]. However, to the knowledge of the authors, there is no technical framework proposed that describes how PDE can be proactively analysed. In addition, the authors have not found any previous work that incorporates threat intelligence as part of the DFR procedures for accurately limiting the volume of digital evidence an investigator needs to examine should an incident occur. To this extent, the next section presents the authors' previous work on efficiently introducing threat intelligence into DFR procedures.

3. The Threat Intelligence Informed Digital Forensic Readiness Reference Model

In their previous work [5,6], the authors proposed a model that aimed to increase the level of the operational digital forensic readiness of organisations, meaning to improve their ability to forensically respond and identify the root cause of information security incidents promptly. The innovative idea of that model (Figure 2) was its ability to analyse and utilise effectively the Indicators of Compromise (IoCs) it collected from various Threat Intelligence Platforms (TIPs) for uncovering patterns of malicious activity that relate to malware instances.

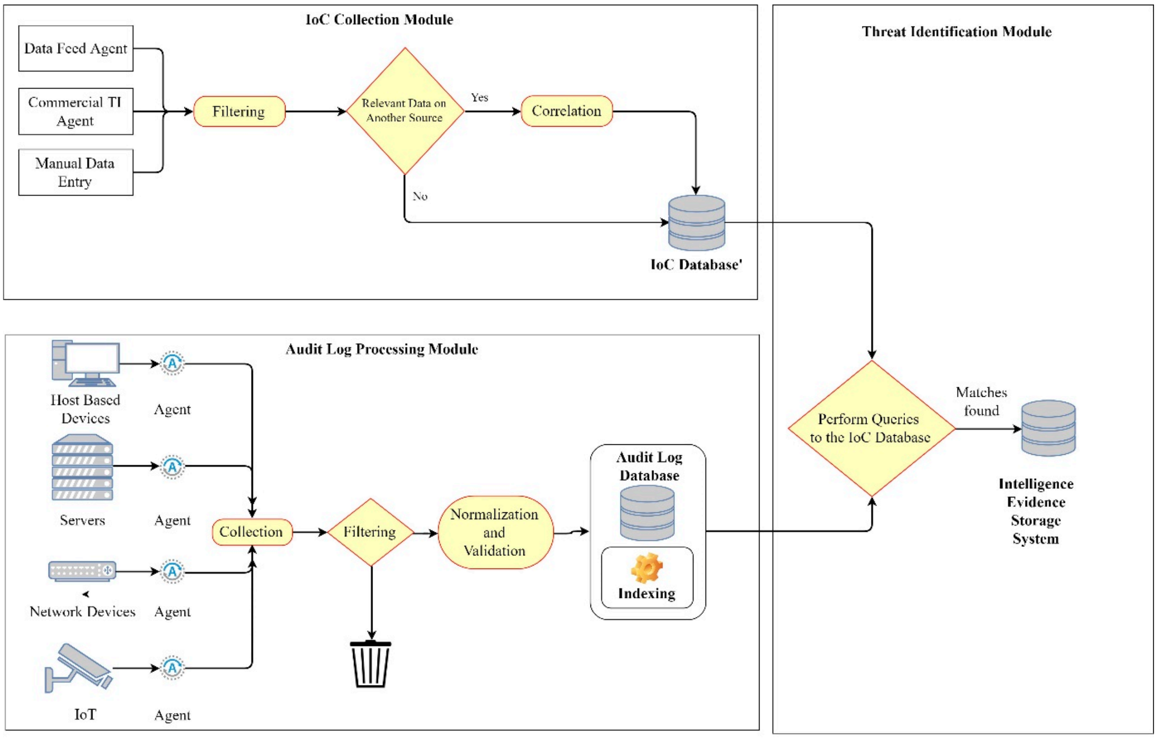


Figure 2. The threat intelligence informed digital forensic readiness process model [6].

The proposed model includes three interrelated, but functionally independent modules whose operation is briefly described in the following subsections.

3.1. The IoC Collection Module

The main role of the IoC collection module is to collect information from various Information Securityrelated sources and transform them into actionable threat intelligence. The introduction of external IoCs is essentially a means for systematically building upon the collective experience and knowledge of past cases, as highlighted by Jusas et al. [24]. To achieve this, the module aggregates data from various TIPs and increases their value through a set of evaluation and correlation procedures. These elaborated data and the relationships between them are stored in a graph database, the Local IoC Database (LIDB). Graph database technology has been selected to support the IoC collection module because it is considered an effective tool for modelling data when a focus on the relationship between entities is a driving force in the design of a data model [25].

Furthermore, the authors acknowledge that data quality is one of the main concerns of collecting information from various TIPs. While any IoC could lead to the discovery of an incident, the systematic generation of high volumes of IoCs can be overwhelming to the analyst, causing adverse effects, which could eventually reduce the efficiency of detection and response. Thus, security teams should consider contextualising the data they collect depending on the threats they possibly face [26].

Influenced by the research of Tounsi and Rais [1] on technical threat intelligence, the authors decided to keep the initial LIDB schema (Figure 3) simple, but functional, while supporting further extensions.

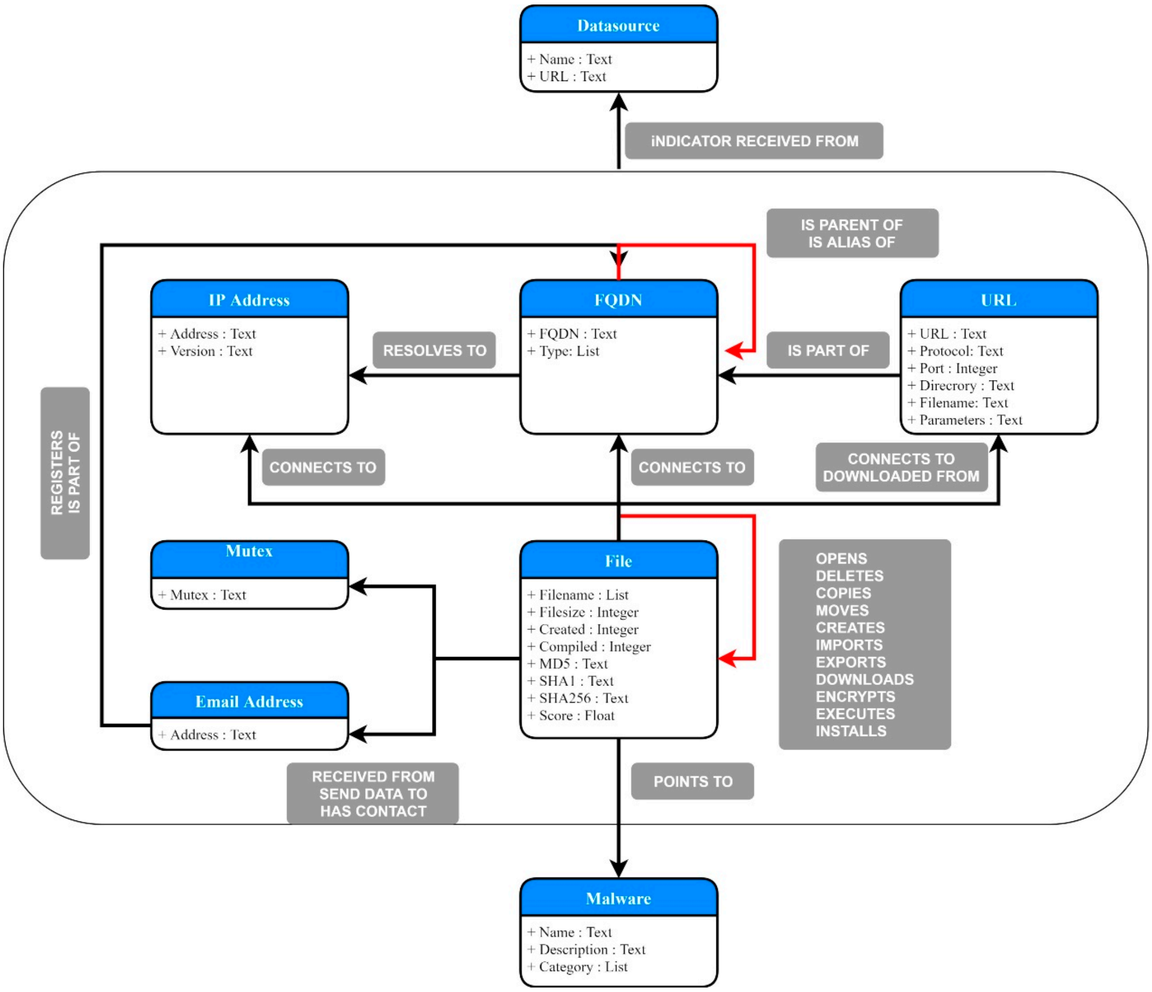


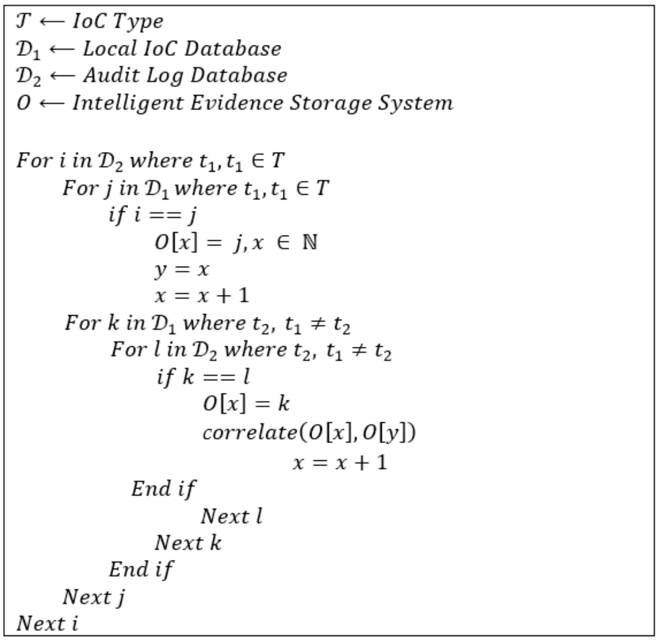
Figure 3. The local IoC database schema [6].

3.2. *The Audit Log Processing Module*

The audit log processing module is an extension of an organisation's existing information security defence mechanisms and aims to gather, normalize, validate, pre-process and forensically store in the Audit Log Database (ALDB) the audit logs produced by the diverse security equipment of the organisation. The inverse indexing features of ALDB support prompt data retrieval requests from the threat identification module, which is explained next.

3.3. *The Threat Identification Module*

The Threat Identification Module (TIM) is the centrepiece element of the threat intelligence informed digital forensic readiness process model, as it is responsible for identifying trails of malicious activity. To this extent, the TIM crossmatches the contents of both the LIDB and ALDB databases, seeks common signs of IoCs, identifies relevant threats and preserves the corresponding evidence in the Intelligence Evidence Storage System (IESS). Thus, the IESS can be considered an efficient inventory of accurate evidentiary records that allows forensic analysts to preview incidents that have the potential of adversely affecting the underlying systems. An effective benefit of the model is its ability to identify, in a timely manner, the attacks that have possibly bypassed an organization's defence measures without raising incident detection alerts. To this extent, the main advantage of the TIM is that it not only identifies suspicious patterns of misuse, but also establishes a prioritised set of root causes that are responsible for provoking adverse effects. Figure 4 depicts a simplified view of the TIM algorithm.



```
$\mathcal{T} \longleftarrow IoC\ Type$
$\mathcal{D}_1 \longleftarrow Local\ IoC\ Database$
$\mathcal{D}_2 \longleftarrow Audit\ Log\ Database$
$O \longleftarrow Intelligent\ Evidence\ Storage\ System$

For i in $\mathcal{D}_2$ where $t_1, t_1 \in T$
    For j in $\mathcal{D}_1$ where $t_1, t_1 \in T$
        if $i == j$
            $O[x] = j, x \in \mathbb{N}$
            $y = x$
            $x = x + 1$
    For k in $\mathcal{D}_1$ where $t_2,\ t_1 \neq t_2$
        For l in $\mathcal{D}_2$ where $t_2,\ t_1 \neq t_2$
            if $k == l$
                $O[x] = k$
                $correlate(O[x], O[y])$
                        $x = x + 1$
          End if
                Next l
            Next k
        End if
    Next j
Next i
```

Figure 4. The algorithm of the threat identification module [6].

It is important to note that unlike supervised modelling, this module and the proposed model as a whole do not intend to detect unknown or zero-day vulnerabilities; rather, they aim to decrease the time and cost of digital forensic investigations by proactively uncovering patterns of malicious activity from stored data logs. Thus, the proposed model aims to offer a lightweight approach to achieving fast and effective triaging of the data, allowing the responder or forensic analyst to filter out quickly the "noise", that is data not related to the system compromise.

4. Evaluation

4.1. Testbed Setup

To conduct the experiments, the authors have set up a virtualized environment with one virtual CPU, 4 GB of RAM memory and the Ubuntu 18.04 × 64 as the guest operating system. On that environment, they installed the Neo4j graph platform [27] and the Elastic Stack Suite [28] to implement the LIDB and ALDB, respectively. The Neo4j database system was selected for the performance it provides while traversing data objects (nodes) and the embedded visualization features [29], while ElasticSearch for its efficient indexing service that allows rapid processing of complex search queries [30].

4.2. IoC Collection Module Operation

As stated earlier, the main role of the IoC collection module is the establishment of a graph database that stores IoCs and the relationships between them. Initially, a set of approximately 1500 malware hash values were imported into the LIDB. Most of this information was obtained randomly from the AlienVault Open Threat eXchange (OTX) repository [31], including approximately 350 hashes that were acquired from the Malware Capture Facility Project [32]. Note that no other types of IoCs were imported into the LIDB at that stage. Each of those hash values was queried against Virustotal [33], AlienVault OTX [31], Hybrid-Analysis [34] and ThreatConnect [35], retrieving additional IoCs and establishing the relationships between them. That process created a graph that contained 93,239 IoCs (file hashes, IP addresses, URLs, Fully Qualified Domain Names (FQDNs), domains and mutexes) in total. These newly-created IoCs, which form the LIDB, are stored in the Neo4j graph database.

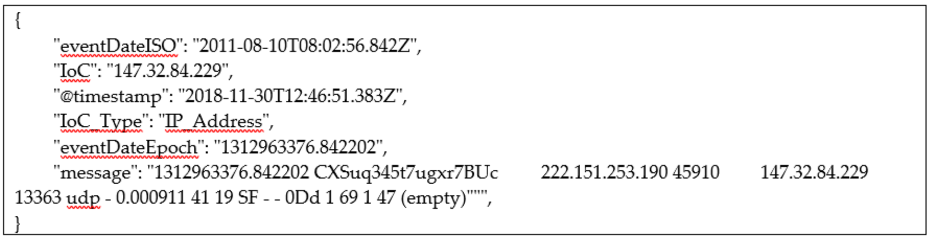
4.3. Audit Log Processing Module Operation

One of the main challenges the authors faced during the experiments was the selection of the test data to feed the audit log processing module. The first option was the in-house production of that data through virtual labs, but this scenario was discarded, as it lacked data credibility. Then, the authors considered that the best approach was the acquirement of external, admissible datasets that could ensure credible experimental results. To this extent, the datasets of the Malware Capture Facility Project were chosen as the most appropriate because they contain de-anonymized text data, where search and regular expression queries for uncovering patterns of possible digital evidence can be applied.

The Malware Capture Facility Project is a sibling of the Stratosphere Intrusion Prevention System (IPS) project. Both of these projects originated from the Czech Technical University of Prague and concentrated on modelling the way real malware instances behave by capturing their network traffic [36,37]. Notably, the credibility and usefulness of the datasets of these projects were acknowledged by Małowidzki et al. [38].

The Malware Capture Facility Project comprises approximately 280 scenarios, each containing network data produced from the execution of a corresponding number of alleged malware instances. Each scenario contains various representations of such data, like pcap files, network flows and audit logs in "Bro" format. Recall that Bro (currently Zeek) is an open-source network intrusion detection system and traffic analyser that can passively monitor network traffic and look for suspicious activity [39,40]. Bro-formatted files have been considered the most appropriate ones for conducting the experiments of this research because they resemble the data commonly seen in operational environments. These files were present in 205 of the 280 scenarios. Thus, only 205 experiments were performed. Despite that limitation, the amount of data used during those experiments exceeded 191 million records and promised credible results. These datasets were used as local log datasets.

The local log data were imported into the ALDB through an automated process that employed the Logstash [28] pipeline, which acted as a file collector, and the ElasticSearch engine [28], which in turn operated both as a NoSQL database and content indexer. The structure of the NoSQL documents of that database contained a limited number of fields that were similar to those of the LIDB. A field storing the contents of every log file in raw format was also added to the NoSQL schema. Such a compact structure can improve search speed and preserve storage space while ensuring the integrity of source data. Figure 5 depicts a sample ALDB entry. The most interesting information from this sample is the IP address, which is the IoC that can be queried against the contents of the LIDB. The "message" field at the end of the sample presents the acquired piece of information in its original format.



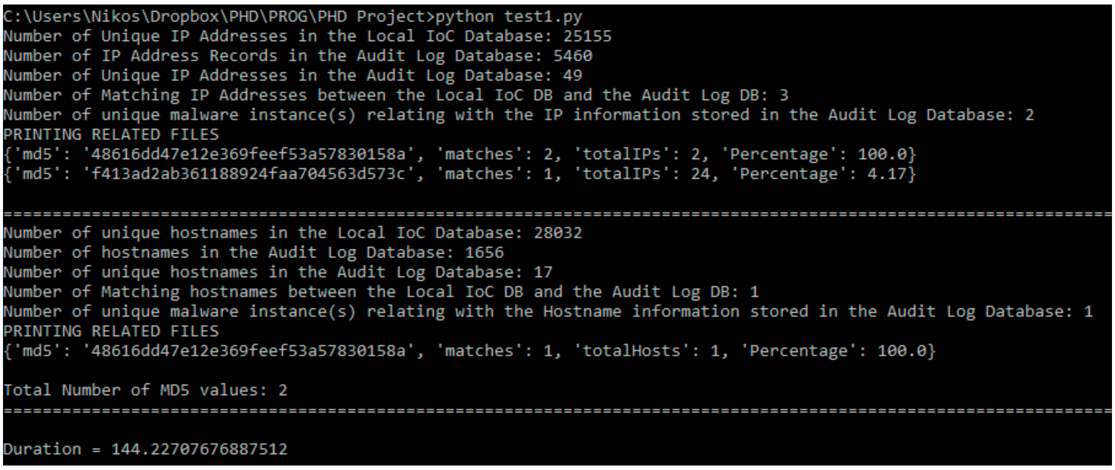
```
{
        "eventDateISO": "2011-08-10T08:02:56.842Z",
        "IoC": "147.32.84.229",
        "@timestamp": "2018-11-30T12:46:51.383Z",
        "IoC_Type": "IP_Address",
        "eventDateEpoch": "1312963376.842202",
        "message": "1312963376.842202 CXSuq345t7ugxr7BUc        222.151.253.190 45910        147.32.84.229
13363 udp - 0.000911 41 19 SF - - 0Dd 1 69 1 47 (empty)""",
}
```

Figure 5. Sample record of the Audit Log Database (ALDB).

4.4. Threat Identification Module Operation

The TIM used the datasets stored in the ALDB database sequentially for conducting the experiments. Each test started with a process that looked for similar IOC fields between the ALDB and LIDB and created a temporary list of similarities. Then, the IoCs of that list looked back at the LIDB graph database and formed a subgraph with common matched IoCs. Considering that some of the IoCs contained malware hash values, it became clear that the TIM managed to identify a list of malware samples that were related to the data of each experiment. The malware outcomes were subject to a simple scoring algorithm that counted each sample's graph connections and thus drew better conclusions about the ones that had possibly affected the target systems.

Figure 6 presents sample results of the operation of the TIM. In this example, the local log file that was imported into the ALDB database contained 5460 records with IP addresses and 1656 records with FQDNs. It did not contain any information about malware hash values. The TIM identified two malware samples as the ones that were possibly responsible for the production of part of the network traffic contained within the log file mentioned earlier. The evaluation of the results proved that TIM correctly identified the malware that was responsible for that event.



```
C:\Users\Nikos\Dropbox\PHD\PROG\PHD Project>python test1.py
Number of Unique IP Addresses in the Local IoC Database: 25155
Number of IP Address Records in the Audit Log Database: 5460
Number of Unique IP Addresses in the Audit Log Database: 49
Number of Matching IP Addresses between the Local IoC DB and the Audit Log DB: 3
Number of unique malware instance(s) relating with the IP information stored in the Audit Log Database: 2
PRINTING RELATED FILES
{'md5': '48616dd47e12e369feef53a57830158a', 'matches': 2, 'totalIPs': 2, 'Percentage': 100.0}
{'md5': 'f413ad2ab361188924faa704563d573c', 'matches': 1, 'totalIPs': 24, 'Percentage': 4.17}

=================================================================================================================
Number of unique hostnames in the Local IoC Database: 28032
Number of hostnames in the Audit Log Database: 1656
Number of unique hostnames in the Audit Log Database: 17
Number of Matching hostnames between the Local IoC DB and the Audit Log DB: 1
Number of unique malware instance(s) relating with the Hostname information stored in the Audit Log Database: 1
PRINTING RELATED FILES
{'md5': '48616dd47e12e369feef53a57830158a', 'matches': 1, 'totalHosts': 1, 'Percentage': 100.0}

Total Number of MD5 values: 2
=================================================================================================================

Duration = 144.22707676887512
```

Figure 6. Sample execution of the threat intelligence module.

5. Results

As previously mentioned, all the datasets used during the experiments contained network data. Therefore, the first challenging task was to identify accurately the malware instances contained in the network data through a series of correlations that existed in the LIDB. The findings of the experiments were verified with the Malware Capture Facility Project since the project's creators provided information about the malware-executable files inside that network traffic. Experiment outputs containing multiple possible outcomes were subject to ranking procedures. The ranking algorithm consulted the LIDB, evaluated their relationships and assigned them a score, highlighting the ones that produced each dataset's contents. The algorithm consulted the LIDB, evaluated IoC relationships and ordered these relationships accordingly, thus pointing out the most likely IoCs included in the dataset.

The experiments revealed that in 176 of 205 experiments, the proposed model accurately identified the malicious files that were related to the corresponding dataset, considering these results as True Positives (TP). Thus, the algorithm initially reached a positive rate of 86.85% and a negative rate of 14.15% (Table 1). In most cases, the results yielded an IoC subset of a relatively small cardinality. Having a limited number of possible IoC candidates to evaluate as malware is highly desirable because it assists digital forensic analysts to find the root cause of an incident easily. Figure 7 shows that only a few experiments returned an increased number of malevolent alternatives; however, in most cases, the scoring algorithm identified the most relevant and prioritised it correctly.

Table 1. Malware instances correctly identified during the experiments.

	Malware Instances Identified	Malware Instances Not Identified
Number of Experiments	176	29
Percentage Rates	85.85%	14.15%

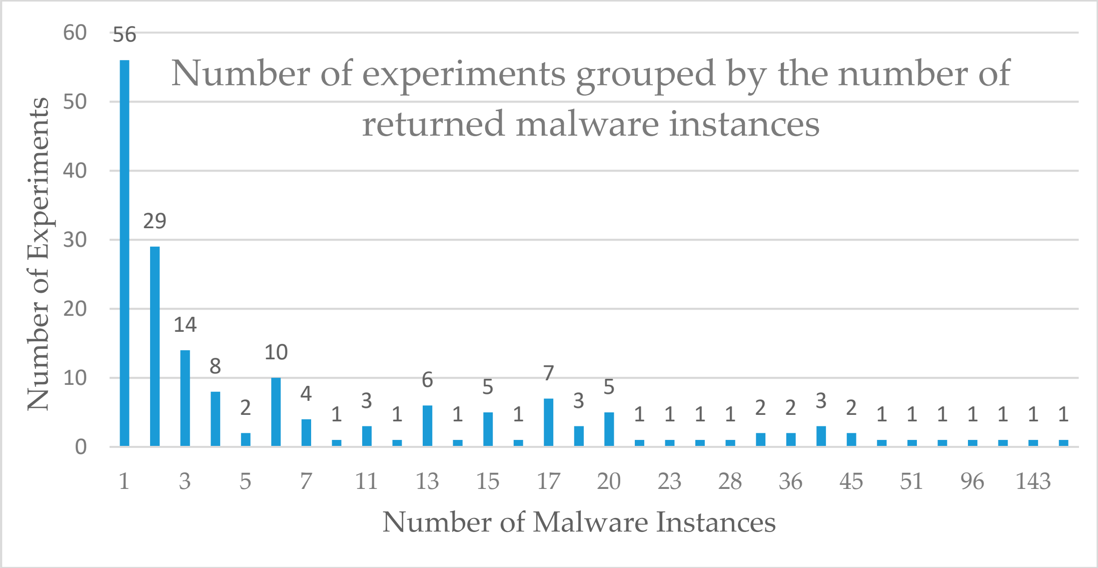


Figure 7. Distribution of the experiments in proportion to the malware instances each experiment returned.

The detection success rate can be considered adequate, as this was the outcome of a lightweight triaging algorithm; hence, the trade-off between computational cost and performance can be acceptable, as presented in the performance analysis below. However, failure to identify the malware instances deserved further investigation since a triage process should, by design, show preference to higher false positives than false negatives.

Acknowledging that time is an essential parameter for evaluating whether a forensic readiness scheme is efficient [7], a performance analysis was conducted. Of all the experiments the authors ran, the most time-consuming one lasted for 251.2 s, which can be considered an acceptable timeframe given the number of records checked and the computing resources used. The average execution time of the experiments was 95.6 s. Having more entries in both LIDB and ALDB databases can slightly increase the investigation time. However, such an increase is trivial compared to traditional digital forensic approaches, which often require a great deal of time and involve processes with high computational costs to reach similar results [41]. Figure 8 shows how the amount of data within the ALDB and the levels of their repeatability shaped the speed of the proposed model.

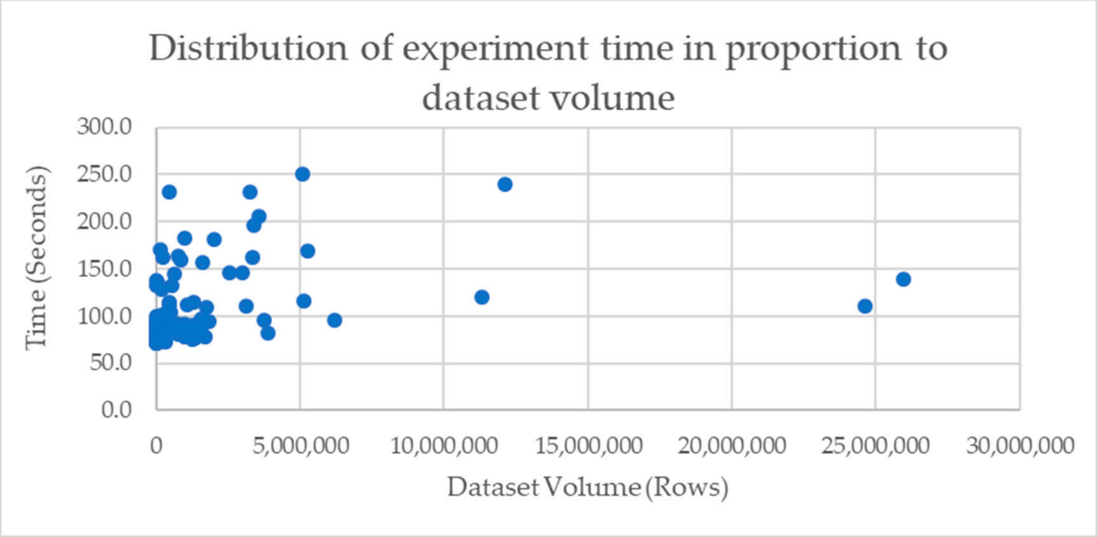


Figure 8. Scatterplot showing how experiments are distributed in proportion to the number of records each dataset contains.

The remaining 29 malware instances that were not correctly identified by the TIM were further examined to establish the reason for the mismatches. The study revealed the following interesting cases:

1. TIPs did not report any malicious network IoCs in 10 of these 29 experiments.
2. The TIM did not return any results (no hash value) in 12 of these 29 experiments.
3. The TIM returned false results (different malware hash value) in seven of these 29 experiments.

The evaluation of the first case indicated that the algorithm behind the TIM did not correctly identify any files as malicious because the malware included in the dataset did not produce any network IoCs, so there was no potential connection with the IoCs in the LIDB. It is more likely that the malware had a different target, for example hosts, and not any network activity. To this extent, these results can be considered as True Negatives (TN).

In the second case, the TIM returned zero results because even though the malware in question did produce potential IoCs, these IoCs were not included and were missing from the ALDB. Thus, these results could be considered as False Negatives (FN).

In the remaining third case, the TIM managed to find correlations between the LIDB and the ALDB; however, it failed to identify the malware instances correctly. The reason for this is that it correctly correlated the produced malware's IoCs to a different malware than the one actually included in the dataset originally, thus identifying a larger family of malware than the original dataset had predicted. This category can be characterized as False Positives (FP).

Despite the misidentification, the correlations among IoCs that exist in the LIDB can reveal additional substantial evidence, such as the presence of droppers, downloaders, backdoors or similar types of malicious activity. Figure 9 presents how the proposed model finally managed to identify the root cause of an incident correctly, despite its initial inability, by employing host-based indicators (mutexes in this case) and their correlation. In this scenario, the proposed model initially tagged that "Malware 2" was the cause of the incident; however, a deeper analysis revealed that "Malware 1" and "Malware 2" belong to the same family. This observation is particularly interesting, as it highlights the importance of studying "near misses".

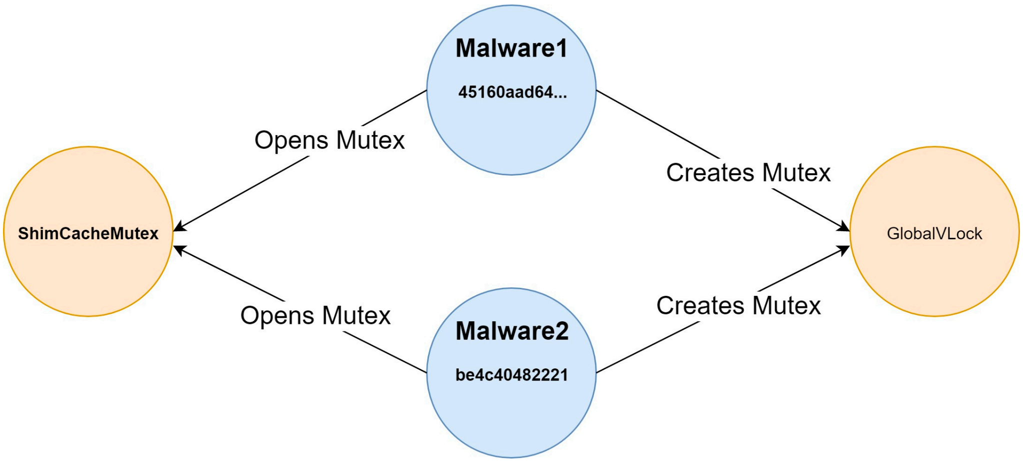


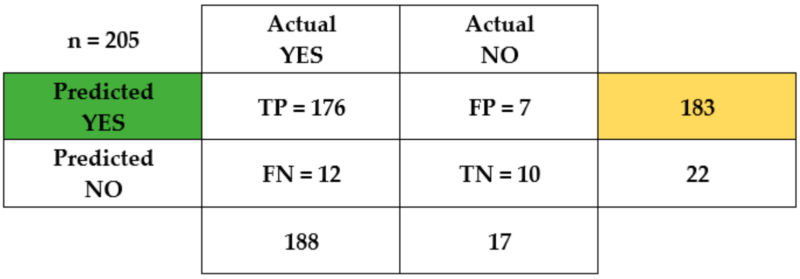
Figure 9. Despite initial false positives, additional analysis often produces correct results.

As stated earlier, the results of the experiments were assigned a category flag (TP, TN, FP, FN), which can be further expressed with the confusion matrix presented in Figure 10. Based on this information, the calculated accuracy, precision and recall, which are given by Equations (1)–(3), respectively, are as follows:

$$\mathcal{A} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{186}{205} = 0.9073 \text{ or } 90.73\% \tag{1}$$

$$\mathcal{P} = \frac{TP}{TP + FP} = \frac{176}{183} = 0.9617 \text{ or } 96.17\% \tag{2}$$

$$\mathcal{R} = \frac{TP}{TP + \ FN} = \frac{176}{188} = 0.9361 \text{ or } 93.61\% \tag{3}$$



n = 205	Actual YES	Actual NO	
Predicted YES	TP = 176	FP = 7	183
Predicted NO	FN = 12	TN = 10	22
	188	17	

Figure 10. Confusion matrix of the algorithm of the Threat Identification Module (TIM).

Another metric indicating the benefits of the proposed approach is the number of audit log entries that are forwarded to the IESS. Experiments showed that trials deriving true positive results (85.85% of all the experiments) decreased the number of the audit log records that were forwarded to the IESS by 90.44%, meaning that fewer than 10% would be considered interesting for further analysis should an incident occur. If false positives, false negatives and true negatives are included in the above calculation, meaning that none of their network data are filtered out, this increases the number of entries that may need additional analysis to 31.38%. Despite such an increase, the volume of data the proposed model managed to save remained high, indicating the usefulness it can provide to forensic analysts.

Finally, it is important to highlight that in data analytics, a system uses training data to define the ground truth [42]. In our approach, however, the ground truth was already known as any collected hash value was considered malicious. Thus, the challenging task of the proposed was is to maximise its true positive results, meaning to identify correctly the presence of a malicious file from any IoCs that may not directly relate to it.

6. Discussion

A novel model leveraging CTI to improve the levels of operational DFR was presented in this paper. To achieve this, the model contained three independent modules, with two of them working proactively and the third used during a forensic triage process. The primary function of the model was the establishment of an inventory containing a limited subset of log records with a high-precision list of cyber-threats, with the aim to minimize the time and consequently the cost of potential digital forensic investigations.

The applicability and effectiveness of the aforementioned model was evaluated through a series of 205 experiments that employed well-known published datasets. In particular, the results indicated that in most cases, the model managed to identify accurately the malware instances that infected the target systems.

In terms of effectiveness and performance, the model managed to limit significantly the network data that may require further analysis by at least 70%. From a digital investigation perspective, this is considered substantial, as it can help digital forensic analysts to examine the contents of audit log files in considerably less time. For example, should we consider that an investigator has to identify critical information in a log file containing n entries, then the time complexity for searching this file is $\mathcal{O}(\mathrm{n})$. Limiting the percentage of entries by κ, where $0 \leq \kappa < 1$ yields a time complexity of $(1 - \kappa) * \mathcal{O}(\mathrm{n})$, signifies that analysis time decreases in proportion to unrelated data. Moreover, it is highlighted that the proactive analysis of the experiments lasted approximately 100 s on average, proving the speediness of the entire process and thus highlighting even more the benefits it can provide to post-mortem analysis.

It should be noted that the above time was achieved due to the architecture of the proposed model, which pre-processed the information it received. Populating both LIDB and ALDB with trustworthy information is potentially time-consuming; however, such a process can run in the background at predefined intervals without interrupting systems' operations and, thus, cannot remarkably affect the overall efficiency of the proposed model.

Last but not least, the authors considered the cost implications of deployment and forensic investigation. In particular, the cost of a forensic analysis project containing several activities can be calculated by Equation (4) [43], where n is the number of activities, t the time to complete the ith activity, v the volume of the ith activity and c the capacity cost of the ith activity. The cost of the capacity refers to the cost of the resources used to perform the activities, such as salaries of employees, equipment and technology costs, rental of office space and any other costs incurred [44].

$$Total\ Activity\ Cost = \sum_{i}^{n} c_i t i v_i \tag{4}$$

Given that the proposed approach did not significantly affect the capacity cost, as it did not require high-end technological resources for the average organisation, it became clear that the total activity cost was proportional to the time and volume of incident investigation, and thus decreased.

In conclusion, the proposed CTI-informed DFR model presented and evaluated in this paper can significantly enhance the DFR state of organizations, as it managed to improve effectively the following KPIs:

- Decrease the volume of information an analyst needs to examine
- Minimise the time of a forensic investigation
- Limit the cost of forensic analysis
- Determine the root cause of an incident in a timely manner and with high precision
- Identify relevant threats that may have affected the security posture of an organization.

7. Conclusions

The challenging nature of the cyber-threat landscape often hampers incident response and digital investigation procedures and thus brings to the forefront the need for efficient DFR strategies. The main motive behind these strategies is their ability to capture accurate information about security incidents in a forensically-sound manner and to assist in identifying the cyber threats that relate to those incidents and mitigating their effects in a timely and efficient manner. To this end, this paper presented and evaluated the authors' previous work on a threat intelligence informed DFR model, which aimed to improve the maturity levels of operational DFR through a set of unified procedures that borrow mechanisms from the fields of threat hunting and security analytics. The model was evaluated through numerous experiments, which proved its operability and efficiency in real-world scenarios. Coupled with the relatively short period of time each experiment needed to be completed, it became clear that the above-mentioned model can be a viable DFR solution. Future approaches could consider further improving the benefits of DFR through the employment of advanced data quality controls of CTI, such as the cyber kill chain- [45] and pyramid of pain-inspired [46] scoring methods. Moreover, the authors are going to employ datasets with host-based logs along with the network datasets already used, as long as such data become publicly available.

Author Contributions: Conceptualization, N.S., V.K. and D.B.; methodology, N.S., V.K. and D.B.; software, N.S.; validation, V.K. and C.I.; formal analysis, N.S. and D.B.; investigation, N.S.; resources, N.S.; data curation, N.S.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, N.S.; project administration, G.P.; funding acquisition, V.K.

Funding: This work has received funding from the European Union's Horizon 2020 research and innovation program under the grant agreement no 830943.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tounsi, W.; Rais, H. A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Comput. Secur.* **2018**, *72*, 212–233. [CrossRef]
2. Serketzis, N.; Katos, V.; Ilioudis, C.; Baltatzis, D.; Pangalos, G.J. Actionable threat intelligence for digital forensics readiness. *Inf. Comput. Secur.* **2019**, *27*, 273–291. [CrossRef]
3. Serketzis, N.; Katos, V.; Ilioudis, C.; Baltatzis, D.; Pangalos, G.J. A Socio-Technical Perspective on Threat Intelligence Informed Digital Forensic Readiness. *Int. J. Syst. Soc.* **2017**, *4*, 57–68. [CrossRef]
4. Bilge, L.; Dumitras, T. Before we knew it: An empirical study of zero-day attacks in the real world. In Proceedings of the 2012 ACM conference on Computer and communications security, Raleigh, NC, USA, 16–18 October 2012.
5. Mandiant. *M-Trends Report*; Mandiant: Alexandria, VA, USA, 2018.
6. Lillis, D.; Becker, B.A.; Sullivan, T.O.; O'Sullivan, T.; Scanlon, M. Current Challenges and Future Research Areas for Digital Forensic Investigation. In Proceedings of the 11th ADFSL Conference on Security and Law (CDFSL 2016), Digital Forensics, Daytona Beach, FL, USA, 13 April 2016.
7. Tan, J. *Forensic Readiness*; Cambridge, MA, USA, 2001. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.644.9645&rep=rep1&type=pdf (accessed on 23 July 2019).
8. Pangalos, G.; Ilioudis, C.; Pagkalos, I. The Importance of Corporate Forensic Readiness in the Information Security Framework. *2010 19th IEEE Int. Workshops Enabling Technol. Infrastruct. Collab. Enterp.* **2010**, 12–16. [CrossRef]
9. International Organization for Standardization. *ISO/IEC 27043: Information Technology, Security Techniques, Incident Investigation Principles and Processes*; International Organization for Standardization: Geneva, Switzerland, 2015.
10. Taylor, C.; Endicott-Popovsky, B.; Frincke, D.A. Specifying digital forensics: A forensics policy approach. *Digit. Investig.* **2007**, *4*, 101–104. [CrossRef]
11. Grance, T.; Hash, J.; Stevens, M. Security Considerations in the Information System Development Life Cycle. *Nist Spec. Publ.* **2004**, *800*, 1.

12. Grobler, C.P.; Louwrens, C.P.; Von Solms, S.H. A framework to guide the implementation of proactive digital forensics in organizations. In Proceedings of the ARES 2010-5th International Conference on Availability, Reliability, and Security, Krakow, Poland, 15–18 February 2010; pp. 677–682.
13. Elyas, M.; Ahmad, A.; Maynard, S.B.; Lonie, A. Digital forensic readiness: Expert perspectives on a theoretical framework. *Comput. Secur.* **2015**, *52*, 70–89. [CrossRef]
14. Al-Mahrouqi, A.; Abdalla, S.; Kechadi, T. Network Forensics Readiness and Security Awareness Framework. Available online: https://researchrepository.ucd.ie/bitstream/10197/6498/1/insight_publication.pdf (accessed on 23 July 2019).
15. Kebande, V.R.; Karie, N.M.; Venter, H.S. A generic Digital Forensic Readiness model for BYOD using honeypot technology. In Proceedings of the 2016 IST-Africa Week, Durban, South Africa, 11–13 May 2016.
16. Kebande, V.R.; Venter, H.S. A Cloud Forensic Readiness Model Using a Botnet as a Service. *Int. Conf. Digit. Secur. Forensics* **2014**, 23–32. Available online: https://www.researchgate.net/profile/Natalie_Walker4/publication/263617788_Proceedings_of_the_International_Conference_on_Digital_Security_and_Forensics_DigitalSec2014/links/0f31753b5cd085c06a000000/Proceedings-of-the-International-Conference-on-Digital-Security-and-Forensics-DigitalSec2014.pdf#page=25 (accessed on 23 July 2019).
17. Kebande, V.R.; Venter, H.S. Obfuscating a Cloud-Based Botnet Towards Digital Forensic Readiness. In Proceedings of the 10th International Conference on Cyber Warfare and Security ICCWS 2015, Kruger National Park, South Africa, 24–25 March 2015.
18. Kebande, V.; Ntsamo, H.S.; Venter, H.S.S. Towards a prototype for Achieving Digital Forensic Readiness in the Cloud using a Distributed NMB Solution. In Proceedings of the European Conference on Cyber Warfare and Security, Munich, Germany, 7–8 November 2016; pp. 369–379.
19. Rowlingson, R. A Ten Step Process for Forensic Readiness. *Int. J. Digit. Evid.* **2004**, *2*, 1–28.
20. Grobler, C.P.; Louwrens, C.P. Digital Forensic Readiness as a Component of Information Security Best Practice. In *IFIP International Information Security Conference*; Springer: Boston, MA, USA, 2007.
21. Valjarevic, A.; Venter, H. Towards a Digital Forensic Readiness Framework for Public Key Infrastructure systems. In Proceedings of the 2011 Information Security for South Africa, Johannesburg, South Africa, 15–17 August 2011.
22. Valjarevic, A.; Venter, H. Implementation guidelines for a harmonised digital forensic investigation readiness process model. In Proceedings of the 2013 Information Security for South Africa, Johannesburg, South Africa, 14–16 August 2013.
23. Elyas, M.; Maynard, S.B.; Ahmad, A.; Lonie, A. Towards A Systemic Framework for Digital Forensic Readiness. *J. Comput. Inf. Syst.* **2014**, *54*, 97–105. [CrossRef]
24. Jusas, V.; Birvinskas, D.; Gahramanov, E. Methods and Tools of Digital Triage in Forensic Context: Survey and Future Directions. *Symmetry* **2017**, *9*, 49. [CrossRef]
25. Miller, J. Graph Database Applications and Concepts with Neo4j. *Proc. 2013 South. Assoc.* **2013**, 141–147. Available online: https://pdfs.semanticscholar.org/322a/6e1f464330751dea2eb6beecac24466322ad.pdf (accessed on 23 July 2019).
26. Bellis, E. The Problem with Your Threat Intelligence 2015. Available online: http://pages.kennasecurity.com/rs/958-PRK-049/images/Kenna_WP_TheProblemwithYourThreatIntelligence.pdf (accessed on 23 July 2019).
27. Neo4j.Neo4j Graph Platform. 2019. Available online: https://neo4j.com/ (accessed on 11 January 2019).
28. Elastic. Elastic Stack Suite. 2018. Available online: https://www.elastic.co/products (accessed on 2 December 2018).
29. Vicknair, C.; Macias, M.; Zhao, Z.; Nan, X. A Comparison of a Graph Database and a Relational Database: A Data Provenance Perspective. Available online: https://john.cs.olemiss.edu/~{}ychen/publications/conference/vicknair_acmse10.pdf (accessed on 23 July 2019).
30. Kalyani, D.; Mehta, D.D. Paper on Searching and Indexing Using Elasticsearch. *Int. J. Eng. Comput. Sci.* **2017**, *6*, 21824–21829. [CrossRef]
31. AlienVault. ‘AlienVault-Open Threat Exchange. 2018. Available online: https://otx.alienvault.com/dashboard/new (accessed on 2 December 2019).
32. Stratpsphere Lab. Datasets Overview—Stratosphere IPS. Available online: https://www.stratosphereips.org/datasets-overview/ (accessed on 4 December 2018).
33. VirusTotal. VirusTotal Malware Analysis Platform. 2018. Available online: https://www.virustotal.com/ (accessed on 2 December 2018).

34. Crowdstrike. Hybrid Analysis-Free Automated Analysis Service', 2018. Available online: https://www.hybrid-analysis.com (accessed on 2 December 2018).
35. ThreatConnect. ThreatConnect Enterprise Threat Intelligence Platform. 2011. Available online: https://threatconnect.com (accessed on 11 January 2019).
36. Garcia, S. Modelling the Network Behaviour of Malware to Block Malicious Patterns. The Stratosphere Project: A Behavioural Ips. Available online: https://www.virusbulletin.com/uploads/pdf/conference/vb2015/Garcia-VB2015.pdf (accessed on 23 July 2019).
37. Stratosphere Lab. Stratosphere Datasets. 2015. Available online: https://www.stratosphereips.org/ (accessed on 10 August 2018).
38. Małowidzki, M.; Berezi, P.; Mazur, M. Network Intrusion Detection: Half a Kingdom for a Good Dataset. ECCWS 2017 PDF. In Proceedings of the 16th European Conference on Cyber Warfare and Security, Dublin, Ireland, 29–30 June 2017.
39. The Zeek Project. Bro Network Intrusion Detection System. 2018. Available online: https://docs.zeek.org/en/latest/intro/index.html (accessed on 3 January 2019).
40. Mehra, P. A brief study and comparison of Snort and Bro Open Source Network Intrusion Detection Systems. *Int. J. Adv. Res. Comput. Commun. Eng.* **2012**, *1*, 383–386.
41. Volonino, L. Electronic Evidence and Computer Forensics. *Commun. Assoc. Inf. Syst.* **2003**, *12*, 1–24. [CrossRef]
42. Friedberg, I.; Skopik, F.; Settanni, G.; Fiedler, R. Combating advanced persistent threats: From network event correlation to incident detection. *Comput. Secur.* **2015**, *48*, 35–57. [CrossRef]
43. Reddy, K.; Venter, H.S.; Olivier, M.S. Using time-driven activity-based costing to manage digital forensic readiness in large organisations. *Inf. Syst. Front.* **2012**, *14*, 1061–1077. [CrossRef]
44. Roberts, K.; Anderson, S.R. *Time-Driven Activity-Based Costing: A Simpler and More Powerful Path to Higher Profits*; Harvard Business Review Press: Brighton, MA, USA, 2007.
45. Lockheed Martin. The Cyber Kill Chain. Available online: https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html (accessed on 10 February 2019).
46. Bianco, D. The Pyramid of Pain. 2013. Available online: http://detect-respond.blogspot.gr/2013/03/the-pyramid-of-pain.html (accessed on 26 September 2017).