University of Sheffield

Cyber Threat Hunting and Digital Forensics Research Report

Jagpreet

A Research Report on Internet-of-Things Forensics and Digital Forensics & Natural Language Processing

Contents

Chapter 1

Part A: IoT Forensics

1.1 Briefly describe what "IoT forensics" is and why the area is important

Kevin Ashton(1999) coined the term Internet of Things(IoT) to describe "a world where the Internet is connected to the Physical World through ubiquitous sensors"[1]. IoT devices allow us to replace human-human or human-computer interactions with machine-to-machine interactions allowing data collection and processing at enormous scales and speeds.

According to Shams Zawoad [2], IoT Forensics is a branch of Digital Forensics combining three different fields, namely:

- Device Level Forensics: It may provide local storage and memory, sensor data.
- Network Forensics: Network Logs and nature of data exchange.
- Cloud Forensics: As Clouds provide the destination for the storage and processing of data, Information such as application data, cookies, session logs etc can be extracted[3].

IoT Forensics consists of the following four stages[4] :

- Identification: Focus is on Device and related infrastructure, for example, router.
- Preservation: Acquiring Data

- Analysis: Analyzing Data using Forensic tools
- Presentation: Detailing and Presenting the findings.

The importance of IoT forensics can be shown by the following:

- IoT devices hold a wide range of Data and sometimes act as the central node for a wide array of other IoT devices, for example, Amazon echo controlling other smart bulbs etc, and can fill gaps or augment investigations[5]
- The traditional Digital Forensics approaches and tools are limited as evidence from these devices have limited visibility and a short survival period[6]. There is a need to develop new tools and frameworks to tackle the challenges.
- The number of IoT devices by 2030 will be around 29 billion[7], and tools and frameworks need to be developed for analysis of these devices.

1.2 Describe which tools are used for IoT forensics?

The IoT system as a whole is more distributed and heterogeneous than traditional targets of Digital Forensics such as Mobiles or Computers, However they also contains parts which may be common in most devices such as storage, memory and processors[8]. Some of the tools used for IoT Devices are:

- Common Digital Forensic Tools:
  - FTK Imager: Forensic Toolkit Imager is an common forensic tool to save an image of the local storage of the device.
  - Volatility: To analyse and perform forensics on the memory of the device.
  - Kismet & Wireshark[9]: Kismet is a passive wireless sniffer, Wireshark is a common network protocol analyzer.
  - ExifTool[9]: By Phil Harvey, To analyse the metadata

  - Autopsy and Encase : To find and present evidence from the device, which can range from session history to type of data collected.
  - DICE[10]: Databse Image Content Explorer to retrieve and access data from RDBMS systems such as MySQL.

- IoT Forensics[11]:
  - ConvertPDML[11] [12]: To capture data between Whatsapp client and server.
  - DRone open-source Parser(DROP)[11]: To analyse DJI drone and correlate DAT files to TXT to link the user of a specific device.
  - Cloud-based IoT Forensic Toolkit[11] [13]:It uses an API to extract data from the Amazon Alexa ecosystem.
  - File System N-View(FSNView)[11]: To study the external storage of Xbox 360.
  - Control Program Logic Change Detector(CPLCD)[11]:Programmable Logic Controllers are used in various Industries, and were also the targets of the Stuxnet Virus. CPLCD can detect attacks on the Siemens PLC such as reprogramming and alteration of memory variables.
  - Volatility plugin for VR[14]: It is a plugin for Votality plugin for VR devices, particularly HTCVive to extract memory artefacts such as location, state and class of devices.

1.3 Describe two significant challenges faced when conducting IoT forensics and explain why they are significant.

The National Institute of Standards and Technology(NIST) identifies 65 challenges associated with cloud and IoT Forensics[15][16].They can be grouped under categories such as identification, collection, preservation, Analysis & Correlation, attack attribution, evidence presentation[15]. Following are the two challenges from which almost all other challenges stem from :

- Distributed and Volatile Data[11]: Data is often separated and either stored locally or remotely in the cloud, As these device seldom possess large storage capacities data stored can have a volatile nature(short lifespan) and can be overwritten easily[4]. As the data can be distributed over a wide network of data centers, It becomes hard to identify the location of data, Furthermore once identified the location of data maybe outside the jurisdiction of the investigating body and subject to different regulations, privacy issues or the access of investigators maybe limited.

  The wide variety and forms in which IoT devices come also further complicates the issue as many devices in the same use-case class may have very different data storage policies and methods. Encryption of data by these devices can also hinder analysis and access to the data. The volatile nature of the evidence may also not allow evidence collection in a "forensically sound manner"[15].

- Diversity of Devices & Protocols[11]: Newer Technologies are coming in every day that change and transform IoT devices in both hardware and software, for example, one generation of Amazon Echo dot could be accessed through Debug Ports to retrieve data, but they were removed in the next generation[8].The range of devices produced by different companies and in different timeframes results in many different software and hardware architectures, different Operating Systems, File structures and Protocols of Connecting to other devices or with the cloud, for example, many different protocols used for connections by IoT devices are BLE(Bluetooth Low Energy), Zigbee(IEEE standard for Personal Area Networks)[11].

  Due to this diversity and continuously changing technologies, holistic frameworks and tools are lacking, the problem that is further exacerbated by lack of funding and interest in the area. Most of the tools available are degraded with change in technology or not reliable enough[11].

1.4 Describe how the two challenges identified in (1.3) are addressed in practice.

One of the proposed solutions that can tackle the above two challenges was given by Nancy Scheidt[17], It is called Hybrid Forensics IOT Server (HFIoTS). It solves many of the problems regarding the location of data, privacy concerns, access to investigation, preservation of data, and diverse internal structures of devices.

HFIoTS is designed to run on hierarchial and distributive levels, and is made and regulated by government. Its strucutre is as follows:

- Every Country has a main HFIoTS server which further has many sub-servers covering the rest of the country.
- The main HFIoTS can be contacted if information needs to be accessed from another country, this allows HFIoTS to be a platform fro investigation purposes nationally and internationally as well supporting the compatibility of all different OS, file structures as information is stored in a format independent of device specific policies.
- When an IoT device connects to an HFIoTs, it is assigned an unique identity(called DNA[17]) and information is stored on the sub-server of the region in which the owner resides. The content of the device is being stored on the server and new device information are updated regularly to ensure HFIoTS is up-to-date.

Hany F. Atlam[18] suggested a Blockchain-Based Investigation Framework[15], where the digital evidence collected and updated will ensure its validity thereby solving the problem of Preserving Data and ensuring integrity.A copy of the ledger could be maintained by all relevant stakeholders such as government, device owner or companies etc to ensure protection against violation of privacy and "Single point of failure"[15].

Other researchers like Yaqoob [6] have argued for "on-the-fly" data processing in place of "store-than-process" approach to tackle the amount of data growth[15]. There has been many development of standards by organisations such as ISO which grant some homogeneity to the

device of similar use-cases, and legislative steps for example,The GDPR legislation warrants companies to report a breach within 72 hours[15], thereby giving Investigators a chance to react and collect evidence before it is destroyed.

Chapter 2

Part B: Digital Forensics and Natural Language Processing.

The research paper chosen for this assignment is "Applying Natural Language Processing for detecting malicious patterns in Android applications" by Shahid Alam[19], available here `https://find.shef.ac.uk/permalink/f/98odl8/TN_cdi_webofscience_primary_000709481500004`.

2.1 Explain what the Digital Forensics problem being addressed is and why it is important.

The Digital Forensics problem being addressed is Malware Analysis of Android applications. Android is the most popular mobile operating system in the world, with around 2.87 million applications on the Google Play Store alone, which makes detecting malicious applications a difficult task. The research paper uses Natural Language Processing to build a binary classifier of an intermediate representation of the application that can automatically classify new Android applications into benign or malicious[19, p 1]thereby automating the process.

Mobiles today are a prime target of Cyber-Criminals, as evidenced by around 40 million malware detections in the 4^{th} quarter of 2020, of which 3 million were new [20]. Accompanied by the sheer quantity of malware are the sophistication and diverse modes of exploitation

used by these malware, such as smishing(sms+phishing), Cryptomining, Game Hacks and Fake Messaging applications[21]. These challenges demand the need for advanced tools and techniques, and automation to tackle the problem[19].

2.2 Summarise how NLP has been applied to the problem.

The application is first converted into an intermediate language called Malware Analysis Intermediate Language(MAIL), which allows the program to be broken down into control flow patterns(sentences) and block patterns(words). An Index of MAIL Programs is made(IMP) by taking all the Control Flow Graphs of all functions in all samples and using Lemmatization to make it amenable to techniques of Natural Language Processing.The final step before apply NLP techniques is to build separate IMP_b (benign) and IMP_m (malicious), and taking out the common features from IMP_b and IMP_m out.

Using the bag-of-words[22] technique, a corpus is built of the tokenized dictionary as follows:

$$IMP_{corpus} = \sum_{k=1}^{S} \sum_{n=1}^{N} IMP_k T_n$$

where S is number of samples, N is number of functions in sample k, and

$$IMP_k T_n = \{id, cf, df\}$$

is the token n in sample k, id is unique integer ids of tokens, cf is collection frequency(frequency of token in sample) and df is document frequency (number of samples containing this token).

As the objective is to compare semantic similarities of programs, The assignment of weights to features must be such that specific parts of a program are assigned more weight than the common ones and tokens more common across samples are given less weight compared to tokens which occur in a few samples, as they contribute to the semantic structure more.This is achieved by using a technique called Term Frequency - Inverse Document

Frequency(TD-IDF)[23]. The weight of each token is computed as follows :

$$W_{i,k} = cf_{i,k} \times log\left(\frac{S}{df_i}\right)$$

where, i is a token in sample k and S is number of samples.

The final step is to Compute a Similarity Index called Similarity Index for MAIL Programs (SIMP). A weighted vector space model is generated using Doc2Vec(which converts whole documents using Word2Vec)to form $SIMP_{corpus}$.

Every new sample's feature vectors are compared to the important malware features in the $SIMP_{corpus}$ to compute a similarity score for each vector using Cosine Similarity[19].The final similarity score is calculated by averaging the scores of all vectors of a sample. If the final is score is above a threshold, the sample is classified as malware and if not, classified as benign.

2.3 Explain how successful the application of NLP was to this problem.

The dataset used by the research paper used a total of 2023 applications where 1023 were malignant, and 1000 benign.The experiment computes a confusion matrix(which helps visualise the performance of a binary classifier) and uses Mathews Correlation coefficient(MCC) to describe the Confusion matrix as a single number.

The experiment can be regarded as successful as it achieves a high detection rate of 93 percent and MCC score of 0.94 signifying a high correlation between predictions and true values.

However, there are few limitations of the above experiment which may affect its success rate -

- The dataset chosen is very small compared to the number of applications and at very best can capture only few types of malware and their patterns.

- As pointed out in the research paper itself it cannot detect Zero-Day Malware or malware that is encrypted, or requires to download additional code.

2.4 Identify one aspect of the work you think should be researched further and briefly outline how this might be done.

The use of an Intermediate representation to analyse applications is an interesting approach that can be adopted in other areas of analysis, It allows platform independent analysis of code, which can lead to greater automation and collaboration for better results in not only malware analysis, but also tackling Piracy, and patent infringement of Software Products.

The Intermediate representation approach can be extended to map out API calls, and network behaviour in the cloud or within organizations to form a monitoring system which can detect threats or malicious behaviour in real time. It can also be used to from a archive of known malicious programs and their targets such as Financial Data or unauthorized access, wherein all the patterns used by the programs to facilitate exploitation can be mapped out so as to allow better design of defensive software such as Anti-Viruses and Firewalls.

Bibliography

[1] Thorsten Kramp, Rob van Kranenburg, and Sebastian Lange. *Introduction to the Internet of Things*, pages 1–10. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-40403-0. doi: 10.1007/978-3-642-40403-0_1. URL https://doi.org/10.1007/978-3-642-40403-0_1.

[2] Shams Zawoad and Ragib Hasan. Faiot : Towards building a forensics aware eco system for the internet of things. 06 2015. doi: 10.1109/SCC.2015.46. URL https://www.researchgate.net/publication/277076879_FAIoT_Towards_Building_a_Forensics_Aware_Eco_System_for_the_Internet_of_Things.

[3] Hany F. Atlam, Ezz El-Din Hemdan, Ahmed Alenezi, Madini O. Alassafi, and Gary B. Wills. Internet of things forensics: A review. *Internet of Things*, 11:100220, 2020. ISSN 2542-6605. doi: https://doi.org/10.1016/j.iot.2020.100220. URL https://www.sciencedirect.com/science/article/pii/S2542660520300536.

[4] Saad Alabdulsalam, Kevin Schaefer, Tahar Kechadi, and Nhien-An Le-Khac. *Internet of Things Forensics – Challenges and a Case Study: 14th IFIP WG 11.9 International Conference, New Delhi, India, January 3-5, 2018, Revised Selected Papers*, pages 35–48. 08 2018. ISBN 978-3-319-99276-1. doi: 10.1007/978-3-319-99277-8_3.

[5] Sasa Mrdovic. *IoT Forensics*, pages 215–229. Springer International Publishing, Cham, 2021. ISBN 978-3-030-10591-4. doi: 10.1007/978-3-030-10591-4_13. URL "https://doi.org/10.1007/978-3-030-10591-4_13".

[6] Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Arif Ahmed, S.M. Ahsan Kazmi, and Choong Seon Hong. Internet of things forensics: Recent advances, taxonomy, requirements, and open challenges. *Future Generation Computer Systems*, 92:265–275, 2019. ISSN 0167-739X. doi: https://doi.org/10.1016/j.future.2018.09.058. URL https://www.sciencedirect.com/science/article/pii/S0167739X18315644.

[7] Lionel Sujay Vailshery. Iot devices forecast. URL https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/.

[8] Shancang Li, Kim-Kwang Raymond Choo, Qindong Sun, William J. Buchanan, and Jiuxin Cao. Iot forensics: Amazon echo as a use case. *IEEE Internet of Things Journal*, 6(4):6487–6497, 2019. doi: 10.1109/JIOT.2019.2906946. URL https://ieeexplore.ieee.org/document/8672776.

[9] dftools. Df tools. URL https://www.dftoolscatalogue.eu/dftc.home.php/.

[10] Database image content explorer: Carving data that does not officially exist. *Digital Investigation*, 18:S97–S107, 2016. ISSN 1742-2876. doi: https://doi.org/10.1016/j.diin.2016.04.015. URL https://www.sciencedirect.com/science/article/pii/S1742287616300500.

[11] T Wu. Digital forensic investigation of iot devices: tools and methods. 2020. URL https://ora.ox.ac.uk/objects/uuid:7e2a4b13-9dfc-4698-884c-26d8c236f074.

[12] F. Karpisek, I. Baggili, and F. Breitinger. Whatsapp network forensics: Decrypting and understanding the whatsapp call signaling messages. *Digital Investigation*, 15:110–118, 2015. ISSN 1742-2876. doi: https://doi.org/10.1016/j.diin.2015.09.002. URL https://www.sciencedirect.com/science/article/pii/S1742287615000985. Special Issue: Big Data and Intelligent Data Analysis.

[13] Hyunji Chung, Jungheum Park, and Sangjin Lee. Digital forensic approaches for amazon alexa ecosystem. *Digital Investigation*, 22:S15–S25, 2017. ISSN 1742-2876. doi: https://

doi.org/10.1016/j.diin.2017.06.010. URL https://www.sciencedirect.com/science/article/pii/S1742287617301974.

[14] Peter Casey, Rebecca Lindsay-Decusati, Ibrahim Baggili, and Frank Breitinger. Inception: Virtual space in memory space in real space – memory forensics of immersive virtual reality with the htc vive. *Digital Investigation*, 29:S13–S21, 2019. ISSN 1742-2876. doi: https://doi.org/10.1016/j.diin.2019.04.007. URL https://www.sciencedirect.com/science/article/pii/S1742287619301562.

[15] Maria Stoyanova, Yannis Nikoloudakis, Spyridon Panagiotakis, Evangelos Pallis, and Evangelos Markakis. A survey on the internet of things (iot) forensics: Challenges, approaches and open issues. *IEEE Communications Surveys Tutorials*, PP:1–1, 01 2020. doi: 10.1109/COMST.2019.2962586.

[16] Martin Herman, Michaela Iorga, Ahsen Salim, Robert Jackson, Mark Hurst, Ross Leo, Richard Lee, Nancy Landreville, Anand Mishra, Yien Wang, and Rodrigo Sardinas. Nist cloud computing forensic science challenges, 2020-08-25 00:08:00 2020. URL https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8006.pdf.

[17] Nancy Scheidt, Mo Adda, Lucas Chateau, and Yasin Emir Kutlu. Forensic tools for iot device investigations in regards to human trafficking. In *2021 IEEE International Conference on Smart Internet of Things (SmartIoT)*, pages 1–7, 2021. doi: 10.1109/SmartIoT52359.2021.00010. URL https://ieeexplore.ieee.org/document/9556201.

[18] Hany Atlam, Ahmed Alenezi, Madini Alassafi, and Gary Wills. Blockchain with internet of things: Benefits, challenges and future directions. *International Journal of Intelligent Systems and Applications*, 10, 06 2018. doi: 10.5815/ijisa.2018.06.05.

[19] Shahid Alam. Applying natural language processing for detecting malicious patterns in android applications. *FORENSIC SCIENCE INTERNATIONAL-DIGITAL INVESTIGATION*, 39:301270, 2021. ISSN 2666-2817.

[20] McAfee. Mobile threat report. . URL https://www.mcafee.com/content/dam/global/infographics/McAfeeMobileThreatReport2021.pdf.

[21] McAfee. The mcafee consumer mobile threat report. . URL https://www.mcafee.com/content/dam/consumer/en-us/docs/reports/rp-mobile-threat-report-feb-2022.pdf.

[22] Zellig S. Harris. Distributional structure. *¡i¿WORD¡/i¿*, 10(2-3):146–162, 1954. doi: 10.1080/00437956.1954.11659520. URL https://doi.org/10.1080/00437956.1954.11659520.

[23] Christopher D Manning. *Introduction to information retrieval*. Syngress Publishing,, 2008.