A Methodology to Create Synthetic Forensic Smartphone Test Data for Research and Education

Patrik Gonçalves[1], Andreas Attenberger[1], and Harald Baier[2]

[1] Central Office for Information Technology in the Security Sector (ZITiS), Research Unit Digital Forensics, Zamdorferstr. 88, 81677 Munich, Germany
{patrik.goncalves,andreas.attenberger}@zitis.bund.de
https://www.zitis.bund.de/

[2] Research Institute Cyber Defence (CODE), Bundeswehr University Munich, Carl-Wery-Straße 22, 81739 Munich, Germany
harald.baier@unibw.de
https://www.unibw.de/code

Abstract. Digital forensics addresses the recovery and investigation of information in digital devices and is an important discipline for aiding forensic experts and courts in solving crimes. To find the forensically relevant data on confiscated devices, digital forensic experts use a wide range of hardware and software tools. Correspondingly, these tools need to be evaluated with realistic test data to ensure that they work as expected and deliver court-proof evidence. The creation of these test data typically is very time- and resource-consuming and researchers often use published datasets, but these are highly biased towards a specific domain or lack complexity. In this work we propose a method to model a high-level semi-automated approach in creating custom digital corpora for digital devices. For generating synthetic forensic datasets a global scenario is defined to specify case-relevant events, which should be found in the final dataset, e.g. writing a message with criminal content. Each event contains temporal information, i.e. a timestamp of when a particular action is performed. From a timeline of events, the method identifies free time slots and automatically creates further case-irrelevant events, e.g. leisure time or shopping. In our use case, we specify the generation of a mobile dataset with synthetic computer-generated forensic irrelevant events.

Keywords: forensic · corpora · mobile · datasets · generator · automatic.

1 Introduction

Mobile forensics is a subset of digital forensics, which includes the forensic investigation of mobile devices, such as smartphones, cell phones, tablets and wearable devices. Digital forensic data from mobile devices for presentation as evidence in court are gaining importance, as these devices are according to Garfinkel et al. [10] “a primary tool of criminals and terrorists” and therefore potentially containing forensically relevant data. An increasing number of smartphone users worldwide with 6.1 billion in 2019 [18] and therefore about 8 out of 11 people worldwide actively operates a smartphone [19] including devices being witnesses in crimes.

For the development of digital forensic software tools, developers need realistic test data to ensure that these tools are properly working while preserving chain-of-custody to ascertain court-proof evidence [2]. These datasets are also essential for educational or academic work, but these persons often do not have access to use datasets from real cases (real data), as privacy laws limit the usage and distribution, e.g. General Data Protection Regulation for the European Union [7]. According to Grajeda et al. [12] only “3.8% of the newly created [datasets] were released” and identify that “researchers prefer not to share their datasets” and they speculate multiple reasons, e.g. copyright and privacy issues. The need of published datasets is also addressed by the authors in [3, 8–10, 15]. This leaves researchers with two options: a) manually creating custom corpora or b) use published corpora. The first option is typically created by manually operating a device, which is highly resource- and time-consuming and therefore in most cases the latter is the preferred option. On the other hand, published corpora might contain biased information or data related to a specific type of crime. A prominent example is the real data Enron Corpus [6], which contains about half a million of published e-mail messages, but the message content is mostly thematically restricted to one company. Exclusively using such dataset in the validation phase of a forensic tool could result in the tool being too specific for general use [21].

Forensic datasets typically contain data according to a predefined scenario, i.e. the underlying story, which describes a timeline of events. These contain forensic artifacts, i.e. bits of forensic information, which are typical to a specific type of crime, e.g. devices used in crimes related to forgery of documents may contain an unusual high number of pictures, PDFs and document files. Therefore, the contents are highly dependent on the type of crime and the story it represents. Data containing potential information for solving a particular case are called forensically relevant information. The remaining information is classified as irrelevant.

Therefore, manual creation of custom datasets mandates a precise knowledge about typical patterns of application usage by criminals pertaining to specific incidences, i.e. knowing typical criminal behavior. Furthermore, detailed knowledge is needed about information and data structures created by mobile applications on the individual devices. Some work has been done on analyzing the behavior of particular apps [14, 13, 4, 1], but the short update cycles and constant technological change makes it a particular challenge.

In our work, we review available digital forensic generators in Section 2 and based from the work found, introduce a methodology to create semi-automated forensic smartphone datasets in Section 3. In Section 4 we present our work done up to date.

2 Related Work

There have been multiple attempts to create synthetic forensic dataset generators. Moch and Freiling created the Forensig^2 [16] framework, which injects forensic artifacts by directly interacting with the storage device. The framework provides full time control and is able to randomize content, while using the full potential of virtual machines. The *3LSPG* framework by Yannikos et al. [20, 21] allows the creation of content based on discrete time Markov chains. Activities (events) are defined as nodes, e.g. writing a message, opening an app, etc. and transitions between two nodes are modeled with conditional probabilities for subsequent activities performed by a (fictional) suspect. While this approach automatically generates a set of randomized predefined events, the user still needs to specify the behavior of single apps. The *EviPlant* framework by Scanlon et al. [17] is a framework optimized for the creation and distribution of datasets with similar content. Their approach was used for educational purposes, where students had to solve different cases. This framework specifies a base image and the creator benefits from only needing to store the "evidence packages" which contains all artifacts and metadata and thus reducing redundant information. A more recent approach was made by the authors Du et al. [5] which automate the generation of forensic datasets by simulating user interactions in their *TraceGen* framework. A similar approach was realized by Göbel et al. [11] in their *hystck* framework, which creates realistic network traffic and app behavior by simulating user interactions. The last two approaches differ from the previous attempts, as it is not necessary to specify file manipulations as user interactions are simulated in context aware environments. Also, the *hystck* allows specification of user interaction models and thus facilitating reusing code. Despite of the various attempts made in the past, none actively address the generation of mobile forensic datasets, e.g. smartphone data.

3 Methodology

The approaches highlighted in Section 2 do not address the generation of scenarios for complex user interactions such as forensic datasets for mobile devices and background knowledge on app mechanics is needed, but this information is often unknown. Therefore, we propose generating forensic datasets by simulating user behavior and interactions, rather than trying to comprehend and to rebuild the background app mechanics, as proposed by [5, 11]. In addition to being more intuitive, this approach also generates the same same app background data, as manually interacting with the device. Executing a set of events and modifying the system time, results in an automatic generation of a synthetic dataset which may contain highly realistic data. In Section 2 we identified the gap of existing image generators exclusively address desktop environments and to the best of our knowledge, we found none for mobile devices and therefore we choose to generate a forensic smartphone dataset with our method. Consequentially we derive the following core research question:

How can we generate a realistic synthetic forensic dataset?

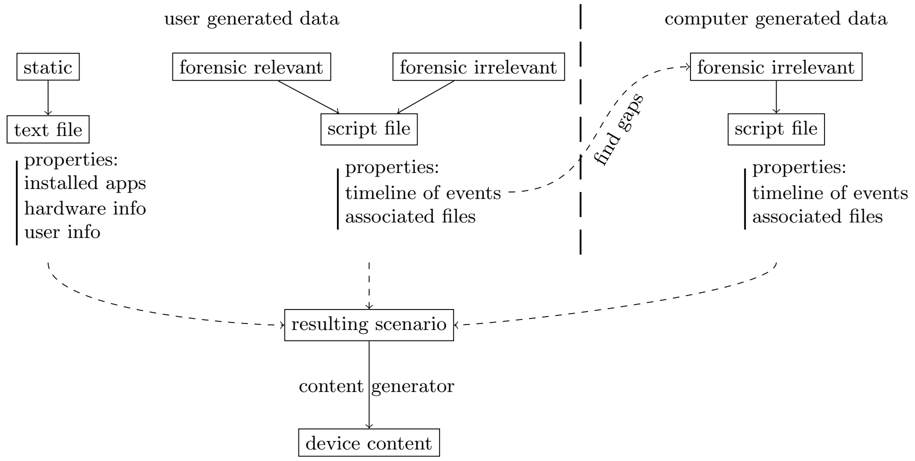
A synthetic dataset in contrast to a real dataset, should not contain any personal data. This means, that the contents and the underlying scenario should not originate from devices, which reference real persons or real crimes. Based from our research question, we argue that a synthetic forensic dataset should meet the following goals: **G1** contain relevant and irrelevant data according to a static script given by a user, **G2** contain forensically irrelevant data which is automatically generated by a system and **G3** contain sufficient amount of data, as found on real devices with similar content.

We propose a methodology, which basically reverses the forensic process [2] and creates data from static reports. The Figure 1 visualizes the structure of our methodology. In detail, we expect a forensic expert to provide a file with forensically relevant and optional irrelevant data, containing a timeline of events. Each event is stored as a human readable text or script file and a timestamp, similar to a forensic report (see goal G1). These events represent the information a forensic expert wants to be included in a synthetic dataset, e.g. a device is used at a particular time to send a message. Specifying a complete realistic set of events is arduous and therefore we plan to automatically generate irrelevant data (see second goal G2), dependent on the timeline of events and static information (e.g. hardware information, user data, installed apps) provided by the forensic expert. Based on this, additional events may be automatically generated, such as activities, which are not relevant for a particular case but also found in real datasets. To achieve this, the system analyzes the events and automatically detects gaps by determining time slots between two subsequent events. From a set of predefined events, the system generates one or multiple events taking place inside this time slot. The resulting events should describe a plausible set of activities and simulating a realistic usage of devices. To determine the amount of data needed and meet our third goal G3, we plan to conduct a survey with forensic experts in law enforcement to determine typical distributions and amounts of data found on confiscated devices. Further, we try to determine typical use patterns of criminals and where experts typically find relevant and irrelevant data.

4 Previous and current work

In previous work we conducted a review for published mobile forensic datasets in scientific publications and web resources. We identified a total of 26 datasets with contents of mobile devices. Further, we analyzed the contents of each dataset with the open-source software *Autopsy* and accounted the occurrences for: audio files, databases, documents with textual information, vector and raster graphic files, video files, user account information, cellular network call logs, contact information on other parties, geospatial points, messenger apps and textual messages. The results of the analysis can be used twofold. First, it can be used as an aid for finding the right dataset for educational and research purposes, without

Fig. 1. The structure of our methodology to generate content on a device. A user provides with static information and forensically relevant and irrelevant data. Based from the temporal information given, the system identifies free timeslots and automatically generates missing irrelevant data. All information is accumulated and then used to populate content on a (virtual) device.



undergoing the need to download and separately analyze each dataset. Second, when creating own datasets one could use the distribution of data found on the datasets to create own synthetic forensic datasets. In the same work, we concluded two main findings: First, we concur that the digital forensics lack of published mobile datasets and more distinctive content is needed. Second, most of the datasets point towards little to no user interactions with short usage times and therefore contain mostly static data, i.e. system and app files and according to Grajeda et al. [12] this would not suffice the critical feature of *quantity*. Currently, we conduct a survey with law enforcement experts to determine typical data distributions found on confiscated storage devices and determine typical applications where experts find relevant and irrelevant data. The results will then be used as a basis for synthetic data with realistic data distributions (see Goal 3). In addition, we investigate an approach to create synthetic geospatiotemporal data. Each event in our method (see Figure 1) is extended by an optional geospatial variable. Correspondingly, the method generates from a set of given spatio-temporal events additional case-irrelevant spatio-temporal events.

References

1. Anglano, C.: Forensic analysis of WhatsApp messenger on android smartphones. Digital Investigation **11**(3), 201–213 (2014)
2. Ayers, R., Brothers, S., Jansen, W.: Guidelines on mobile device forensics. Tech. rep., National Institute of Standards and Technology (2014)

3. Beebe, N.: Digital forensic research: The good, the bad and the unaddressed. In: IFIP International Conference on Digital Forensics. pp. 17–36. Springer (2009)
4. Cahyani, N.D.W., Rahman, N.H.A., Glisson, W.B., Choo, K.K.R.: The role of mobile forensics in terrorism investigations involving the use of cloud storage service and communication apps. Mobile Networks and Applications **22**(2), 240–254 (2016)
5. Du, X., Hargreaves, C., Sheppard, J., Scanlon, M.: TraceGen: User Activity Emulation for Digital Forensic Test Image Generation. Forensic Science International: Digital Investigation (2021)
6. Enron Email Corpus: Enron email dataset. http://www.cs.cmu.edu/ enron/ (2015), accessed: 2021-06-04
7. European Union: General data protection regulation. Official Journal of the European Union **Volume 59**(L 119), 1–88 (2016)
8. Garfinkel, S.: Forensic corpora: a challenge for forensic research. Electron. Evid. Inf. Cent. 1e10 (2007)
9. Garfinkel, S., Farrell, P., Roussev, V., Dinolt, G.: Bringing science to digital forensics with standardized forensic corpora. digital investigation **6**, S2–S11 (2009)
10. Garfinkel, S.L.: Digital forensics research: The next 10 years. digital investigation **7**, S64–S73 (2010)
11. Göbel, T., Schäfer, T., Hachenberger, J., Türr, J., Baier, H.: A novel approach for generating synthetic datasets for digital forensics. In: IFIP International Conference on Digital Forensics. pp. 73–93. Springer (2020)
12. Grajeda, C., Breitinger, F., Baggili, I.: Availability of datasets for digital forensics–and what is missing. Digital Investigation **22**, S94–S105 (2017)
13. Hassenfeldt, C., Baig, S., Baggili, I., Zhang, X.: Map my murder. In: Proceedings of the 14th International Conference on Availability, Reliability and Security. ACM (2019)
14. Levinson, A., Stackpole, B., Johnson, D.: Third party application forensics on apple mobile devices. In: 2011 44th Hawaii International Conference on System Sciences. IEEE (2011)
15. Luciano, L., Baggili, I., Topor, M., Casey, P., Breitinger, F.: Digital forensics in the next five years. In: Proceedings of the 13th International Conference on Availability, Reliability and Security. ACM (2018)
16. Moch, C., Freiling, F.C.: The forensic image generator generator (forensig2). In: 2009 Fifth International Conference on IT Security Incident Management and IT Forensics. pp. 78–93. IEEE (2009)
17. Scanlon, M., Du, X., Lillis, D.: Eviplant: An efficient digital forensic challenge creation, manipulation and distribution solution. Digital Investigation **20**, S29–S36 (2017)
18. Statista: Number of smartphone users worldwide from 2016 to 2023. https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/, accessed: 2021-05-21
19. United Nations: World Population Prospects 2019: Highlights (ST/ESA/SER.A/423). UNITED NATIONS PUBN (2019)
20. Yannikos, Y., Franke, F., Winter, C., Schneider, M.: 3lspg: Forensic tool evaluation by three layer stochastic process-based generation of data. In: International Workshop on Computational Forensics. pp. 200–211. Springer (2011)
21. Yannikos, Y., Graner, L., Steinebach, M., Winter, C.: Data corpora for digital forensics education and research. In: IFIP International Conference on Digital Forensics. pp. 309–325. Springer (2014)