Revisiting the dataset Gap Problem - On Availability, Assessment, and Perspective of Mobile Forensic Corpora

By:
Patrik Gonçalves, Klara Dolos, Michelle Stebner, Andreas Attenberger
and Harald Baier

From the proceedings of
The Digital Forensic Research Conference
DFRWS APAC 2022
Sept 28-30, 2022

Forensic Science International: Digital Investigation 43 (2022) 301439

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS 2022 APAC - Proceedings of the Second Annual DFRWS APAC

Revisiting the dataset gap problem – On availability, assessment and perspective of mobile forensic corpora

Patrik Gonçalves [a, *], Klara Dološ [a], Michelle Stebner [a], Andreas Attenberger [a], Harald Baier [b]

[a] Zentrale Stelle für Informationstechnik im Sicherheitsbereich, Zamdorfer Straße 88, München, Germany
[b] Research Institute Cyber Defence (CODE), Bundeswehr Universität München, Carl-Wery-Straße 22, München, Germany

ARTICLE INFO

Article history:

Keywords:
Digital forensics corpora
Mobile forensics
Availability
Analysis
Quality
Quantity
Timeliness
Dataset
Smartphone
Content
Statistics
Review

ABSTRACT

Digital forensic corpora are essential for education, academic research, tool development and testing. Due to the increasing pervasiveness of mobile devices like smartphones or tablets, the need for mobile forensic datasets is growing, too. However, publications in the IT forensic community show that there is a large gap in publicly available datasets. In this work we focus on mobile digital forensic corpora as one of the main fields of missing datasets and aim at shifting the focus of the digital forensic community on this topic. In order to do so, we provide three main contributions. We first perform a structured search for mobile forensic corpora and show that 31 publicly available mobile forensic corpora exist, 9 of them more than 5 years old and 18 of them more than 3 years old. Second, we assess these datasets with respect to its content compared to an ordinary real mobile image and conclude that most of the 31 datasets contain too few traces to be considered as realistic. Finally, we propose how to proceed to solve the presumable problem of missing mobile forensic datasets.

1. Introduction

The increasing daily use of IT systems and smart devices like computers, mobile phones, smartwatches, and Internet of Things devices, leads to a high demand for digital forensic experts to integrate these devices in their digital forensic investigations. Furthermore, the amount of data to deconstruct has achieved a quantity that a digital forensic analysis can usually no longer be carried out manually. In order to cope with the large amount of data, software and hardware forensic tools are needed that automate parts of the investigation or even the entire analysis. Additionally, the high number of different device models with different builds and operating systems complicates the use of digital forensic tools.

National and international programs such as the *Computer Forensics Tool Testing Program* (CFTT) by the National Institute of Standards and Technology (NIST) provide methodologies and test data for ensuring the reliability of forensic tools among other topics. The use of sophisticated datasets provides law enforcement the ability to make informed choices about acquiring, using and understanding the tools capabilities. Further, tool testing is used to develop novel forensic tools, to improve existing ones or to ensure that available tools consistently produce accurate and objective test results.

In this paper we aim at revisiting the *availability problem of datasets for digital forensics*. We tie in with the previous main publications of Garfinkel et al. (2009) and Grajeda et al. (2017). Our overall goal is to transfer, update, and extend their suggestions and results on digital forensic corpora to the scope of mobile devices and hence shift the focus of the digital forensic community to the still unsolved problem of missing *mobile* forensic datasets. Previous works provided mainly with limited information on current (mobile) forensic corpora, i.e. source location, storage size and used device(s).

In all, we provide three contributions in this paper. We first address the availability aspect of mobile forensic corpora by performing a structured search for and review of mobile forensic

* Corresponding author.
E-mail addresses: patrik.goncalves@zitis.bund.de (P. Gonçalves), klara.dolos@zitis.bund.de (K. Dološ), michelle.stebner@zitis.bund.de (M. Stebner), andreas.attenberger@zitis.bund.de (A. Attenberger), harald.baier@unibw.de (H. Baier).

https://doi.org/10.1016/j.fsidi.2022.301439

datasets. We show that 31 publicly available mobile forensic corpora exist, however 18 of them are more than 3 years old and hence outdated. This again proves the existence of the mobile forensic dataset gap.

Second, we assess each dataset with respect to its content compared to an ordinary real mobile device dataset. We rate each corpus with respect to the assessment categories *quantity* and *quality* as proposed by Grajeda et al. (2017). We conclude that only about every fourth of the 31 datasets contain sufficient traces to be considered as realistic (i.e. their contents are rich in quality and quantity). Additionally, we appraise each dataset with respect to its timeliness where evaluation reveals that more than half of the published mobile datasets is outdated and may not provide precise information for contemporary digital forensic tools to cope with contemporary mobile devices. This again underlines the missing dataset problem. We argue that using the metrics on quantity, quality and timeliness might ease the comparability between different datasets.

Our third contribution is our proposal on how to proceed to solve the presumable problem of missing mobile forensic datasets. We appeal to the forensic community to shift the research focus on creating dataset synthesis frameworks and suggest further necessary steps to come up with a configurable, useable synthesis framework for mobile forensic datasets.

The rest of this paper is organized as follows. After the introduction to the topic in Section 1 we present related work to our paper in Section 2. Then Section 3 shows our systematic review of current public mobile and smartphone corpora and the subsequent analysis with respect to their content. In Section 4 we assess each published mobile forensic corpus with respect to evaluation criteria like quality, quantity, and timeliness. Finally, we present our appeal in favor of dataset synthesis frameworks in Section 5 to overcome the *missing dataset problem* and conclude our work in Section 6.

2. Related work

In this section we review related work in the scope of datasets with a focus on publicly available digital forensic corpora, which are often manually generated and synthetic approaches by presenting some available (mobile) forensic dataset generators.

2.1. Digital forensic corpora

The starting point of addressing the importance of general digital forensic corpora for training and validation purposes is roughly 20 years ago, when Carrier (2005) published tool testing disc images on a *sourceforge* hosted website.[1] Then in 2007 Garfinkel (2007) state the challenge of missing publicly available datasets for the forensic community research. Two years later, again Garfinkel et al. (2009) specify more information on the dataset gap in the forensic community. Based on their DFRWS USA paper, Garfinkel et al. (2009) initiated a website to download digital forensic corpora[2]. However, due to the age of this work, modern mobile forensic datasets were not initially considered. Furthermore, all images were manually crafted.

The missing dataset problem was discussed in the community by overview articles, too. For instance, within publications about general open problems in digital forensics (Beebe (2009); Garfinkel (2010); Lillis et al. (2016)) state along other issues, the problem of missing digital forensic corpora. An important issue with published datasets is the missing ground truth. Having insights into the actual traces is rather important as tool testing and education requires labels to get authoritative results. However, the dataset gap is not specific for the digital forensics community, as we will show in our work. For instance, Abt and Baier (2014) state in the scope of network forensics that published network datasets are rare, anonymized and unlabeled, too.

The most recent milestone is a work published by Grajeda et al. (2017). According to their work, only “3.8% of the newly created [datasets] were released”. The authors list only three sources of mobile forensic corpora. Further, they provide only limited information about the datasets’ content, instead they give a rather coarse overview of the available digital corpora and whether they were generated by real users or were results from experiments. As a result of their work, the authors Grajeda et al. (2017) created a website with a list of available datasets for Cyber Forensics.[3] However, no further information about mobile corpora or about dataset labels are given.

In order to get an impression of the actual digital forensic dataset gap, Luciano et al. (2018) conducted a workshop with forensic experts to determine current challenges and gaps in digital forensics. One key result is that the digital forensic community still lacks of having published and contemporary digital corpora.

One final source for websites to download digital corpora is due to Aman Hardikar,[4] which provides visual mind maps for various topics. It addresses among other topics also Forensic Challenges, where they provide various paths to websites containing digital corpora in general. However, they provide no further details about mobile corpora or labeled mobile content.

As shown before, the reality is that scientists and companies typically create their own datasets to fill the dataset gap. These datasets are often created as outputs from experiments or are part of so called *Forensic Challenges* Grajeda et al. (2017). The *Forensic Challenges* typically created for digital forensic research, provide with relatively complex contents Garfinkel et al. (2009), but often lack a ground truth, because the answers are often not published, but shared only with individuals Woods et al. (2011) or educational and government institutions Digital Corpora (2022) and thus limiting the accessibility to the public. These challenges result often from sophisticated working groups, who created complex scenarios to be solved by the forensic community Woods et al. (2011).

2.2. Image generators

Image generators provide the opportunity to generate a *labeled* forensic image of a persistent storage device (e.g., SSD, USB stick). The digital forensic community created various approaches (Moch and Freiling (2009); Visti et al. (2015); Scanlon et al. (2017); Du et al. (2021); Göbel et al. (2020); Göbel et al. (2022)), however, none of them is currently able to generate mobile images.

One recent approach by Delgado et al. (2021) introduced a proof-of-concept, where they manipulate in using different techniques to modify the database entries in rooted Android emulators. Although, they showed, that all manually injected forensic traces can be recognized by the open-source tool suite Autopsy, they do not provide means to automatically generate larger forensic corpora or creating complex scenarios.

3. Availability and analysis of published mobile forensic datasets

Apart from the scientific related work presented in Section 2,

[1] http://dftt.sourceforge.net/, accessed 2022-05-15.

[2] https://digitalcorpora.org/, accessed 2022-05-15.

[3] https://datasets.fbreitinger.de/, accessed 2022-05-15.

[4] http://www.amanhardikar.com/mindmaps/ForensicChallenges.html, accessed 2022-05-15.

there is no accumulated list of mobile datasets with information about the contents of the respective dataset (, e.g., how many and which sort of files are contained in the dataset and if there are deleted files or fragments of wiped content). Hence, a digital forensic expert is currently not able to decide if a published mobile forensic dataset is of use for his use case or not.

In this section we aim at filling this gap. In Section 3.1 we present the results of our review for published mobile forensic corpora. In all, we found 31 publicly available mobile forensic datasets as shown in Table 1. Next, in Section 3.2 we analyze the content of each of the 31 datasets with respect to basic aspects. For instance, in case of an extracted file we analyze to which *file class* it belongs to (sample file classes are audio files, videos, pictures, textual documents), but we do not take care of the actual file type encoding (, e.g. if a picture is stored as jpg, png, tiff). We furthermore screened for typical *forensic relevant mobile data structures*, e.g., accounts, contacts, geospatial information, which may be stored within files of different file classes. The key result of our analysis is shown in Table 2. We conclude that the currently available corpora contain too few traces to be of utility for tool testing, training or education.

3.1. Availability of published mobile forensic datasets

In this section we present the methodology and the results of our structured (online) search for digital mobile forensic corpora.

3.1.1. Methodology

We subdivide our search into a screening of public web resources and published peer-reviewed papers containing references to forensic mobile corpora, respectively. Our review for public web resources is based on the most commonly used Internet search engines *Google Search*, *Microsoft Bing* and *Yahoo! Search*. We focus on web resources containing any data with smartphone or mobile phone content. The search results range from a simple logical copy of a phone's file system, up to bitwise copies of a phone's complete internal and external storage. From the first 100 search results, we recursively screened the website content. The recursive search first involved a manual retrieval of the search items above. Where applicable we follow linked paths to potential further resource locations. We stop our recursive search when reaching a depth level of three, where level zero is the search engine website. To limit our search, we used in total 30 search combinations, summarized by the following single regular expression:

(digital|smartphone|mobile)\s(forensic\s)?(corpora|test data|dataset|image |CTF)

Similar to the web resource search, we looked for scientific publications since 2000, which deal with information about smartphone and mobile corpora. If additionally a scientific paper contains a download link we extract it. We make use of both scientific search engines and the databases of the four large IT publishers: *Google Scholar*, *ResearchGate*, *IEEE Xplore*, *ACM Digital Library*, *Elsevier Science Direct* and *Springer Link*. In contrast to the general web resource scan, a search in these databases has the advantage to filter out most non-peer reviewed content. We apply the same search pattern as in the web resource search and correspondingly applied a recursive search on the references. In the same manner, we repeated the recursive search up to a recursive depth level of three.

Table 1
An overview of mobile forensic corpora derived by a systematic search on the Internet and scientific literature and including the source, device name, operating system (OS), image extraction type, year of creation, ratings on the datasets' contents (rich:+, neutral:+/−, poor:-) regarding the quantity and quality.

ID	source	device	OS	image type	year	quantity	quality	comment
1	CalPoly CCI	Samsung Galaxy J3	Android	physical	2018	-	+/−	UFED format only
2	DFRWS FC	Motorola Milestone	Android	physical	2009	-	-	failed to reconstruct file system (FS)
3		Motorola Droid	Android	physical	2011	-	-	
4		HTC S620	Windows Phone	physical	2009	-	-	
5	Digital Corpora	Nexus S	Android	physical	2012	-	-	
6		Apple Iphone 3G	iOS	physical	2012	-	-	no iLEAPP support
7		LG Nexus 5	Android	physical	2017	-	-	corrupted timestamps on SD card
8		Google Pixel 3	Android	logical	2020	+/−	+/−	
9		LG Nexus 5X	Android	physical	2018	-	+/−	
10		LG Nexus 5X	Android	physical	2019	+/−	+/−	
11		Google Pixel 3	Android	logical	2019	-	+/−	
12		Google Pixel 3	Android	physical	2021	+	+	
13		iPhone SE	iOS	logical	2020	+	+	running iOS 13.3
14		iPhone SE	iOS	logical	2020	+	+	running iOS 13.4
15		Nexus One	Android	logical	2011	-	-	not many data
16		Nexus S	Android	logical	2011	-	-	not many data
17	NIST CFReDS	iPhone 3 GS	iOS	logical	2012	-	-	corrupted timestamps
18		LG Optimus	Android	JTAG	2012	-	-	failed to reconstruct FS
19		Samsung Galaxy S4	Android	chip-off	2018	-	+/−	
20		HTC Desire 626s	Android	chip-off	2018	-	+/−	
21		LG K7	Android	chip-off	2019	-	+/−	
22		Motorola Moto-E (2G)	Android	chip-off	2018	-	-	failed to reconstruct FS
23		ZTE ZMax	Android	chip-off	2019	-	+/−	
24		HTC Desire S	Android	chip-off	2018	-	+/−	
25		Samsung Galaxy S2	Android	chip-off	2018	-	+/−	
26		HTC One XL	Android	chip-off	2018	-	+/−	
27		HTC One Mini	Android	chip-off	2019	-	+/−	
28	Magnet CTF	Google Pixel 3	Android	physical	2020	-	+/−	
29	Cellebrite CTF	Samsung Galaxy Note 10	Android	file system	2021	+	+	UFED format only
30		Apple iPhone X	iOS	file system	2021	+	+	UFED format only
31		Apple iPhone X	iOS	file system	2021	+/−	+/−	UFED format only

Table 2
Statistical parameters mean, median, standard deviation (SD), minimum and maximum value of the data categories: accounts (Acc), contacts (Con), messenger apps (Msgr), messages (Msgs), calls, geospatial data (Geo), databases (DB), pictures (Pic), video files (Vid), audio files (Aud) and documents (Doc).

	Acc	Con	Msgr	Msgs	Calls	Geo	DB	Pic	Vid	Aud	Doc
mean	15	13	9	94	10	1584	610	9804	1399	555	3399
median	6	8	7	19	0	4	324	4538	24	236	872
SD	22	18	9	154	17	5778	596	13006	6921	992	5842
min value	0	0	0	0	0	0	1	16	0	3	4
max value	114	82	30	605	68	28644	2055	58755	39253	5157	24809

3.1.2. *Filtering*

We first have to remark that the results of both search methods may contain duplicates after the initial screening, however, in a post-processing step we consolidate the results. Secondly, the search results may contain forensic corpora of other devices (, e.g. desktop environments). Again, in our post-processing step we only consider a digital forensic corpus, if at least one mobile forensic dataset is included. Hence, our consolidated list of both search results produces a complete representation of currently available open access mobile corpora.

For the sake of completion, the search for forensic *Capture the Flag* (CTF) datasets revealed many results, but we did not consider those datasets, which did not included a partial or complete image of a mobile device. The CTF challenges primarily created for the computer security hacker community, consist of finding mostly hidden information (also called the ‘flag’) embedded in single files, archives or custom-built apps and rarely contain realistic scenarios as found in the daily work by forensic investigators Trail of Bits (2022).

3.1.3. *Results*

We found in total 5 main sources for freely available mobile corpora on the following websites: *CalPoly CCI*[5] with one forensic mobile dataset for training, *DFRWS Forensic Challenges*[6] with annual forensic challenges containing three mobile devices, *Digital Corpora*[7] with various forensic corpora and challenges containing in total 12 mobile devices, the *NIST CFReDS Project*[8] as a good source for forensic corpora containing 11 mobile devices and CTF events of Cellebrite and Magnet Forensics containing four mobile datasets. Although, we could not find an official link for the CTF dataset from the Magnet User Summit, we were able to find a Google Drive link on a GitBook hosted website.[9]

We not only managed to reproduce the sources containing mobile datasets as found in the work by Grajeda et al. (2017), but we also managed to find additional sources of mobile corpora. This includes the one mobile dataset from the CalPoly CCI, one Android dataset from the Digital Corpora website and four CTF mobile datasets released by Cellebrite and Magnet Forensics. In total, 31 datasets are included in the analysis and assessment in the following sections. Further details on each dataset like publication year, data extraction method, device's operating system are given in Table 1, which we later discuss in the scope of our assessment of the corpora.

3.2. *Analysis of mobile corpora*

In this section we turn to our analysis of all publicly available mobile forensic datasets found in Section 3.1. Our analysis goal is to give an overview about traces and content within each of the gathered dataset. In order to do so, we first introduce in this section categories of data, which we expect on a daily used mobile device. The detailed result of our analysis is listed in Table 3. Based on Table 3, we present our analysis discussion on the gathered corpora in this section and additionally analyze the statistical parameters of the available datasets, e.g. the mean and the standard deviation of the data within each category. Furthermore, our analysis prepares our assessment of the available mobile forensic corpora, which we present in Section 4. More details on each dataset (, e.g. the publication year, technical details) is given in Table 1, which includes evaluation categories, too, and hence is integrated in the assessment section of this paper.

3.2.1. *Analysis categories*

The aim of our analysis is to scan the published mobile forensic corpora for traces and content, which we expect on a daily used mobile device. We therefore categorize the data within a mobile dataset as follows. First, in case of an extracted file we analyze to which *file class* it belongs to. We make use of the following file classes:

- *audio files* (files with audio information, e.g. voice messages)
- *databases* (files with structured data, e.g. SQLite files)
- *(textual) documents* (files with textual information, e.g. notes, PDFs, Word documents)
- *pictures* (vector and raster graphic files)
- *video files* (files with moving visual content)

We point out that we do not take care of the actual file type encoding. For instance, a picture class file may be encoded as bmp, jpg, png, tiff. Second, we screened each mobile forensic dataset for typical *forensic relevant mobile data structures*, which may be stored within files of different file classes. We make use of the following data structure categories:

- *accounts* (credential information on user accounts)
- *call logs* (calls over a cellular network)
- *contacts* (contact information on third parties, e.g. phone book, chat logs and e-mail addresses)
- *geospatial points* (geographical positions with temporal information, e.g. found inside EXIF tags of pictures or stored in cloud services and app databases)
- installed *messenger apps* (medium supporting at least textual communication between two or more participants)
- *messages* (sent/received messages with textual content, e.g. SMS, MMS, chat contents)

3.2.1. *Analysis environment*

In order to ensure repeatability of our work, we use the open source and freely available forensic tool suite Autopsy Carrier (2022) for our analysis in version 4.19.1 on a Windows 10 Pro

[5] https://cci.calpoly.edu/2019-digital-forensics-downloads, accessed 2022-05-15.
[6] http://old.dfrws.org/, accessed 2022-05-15.
[7] https://digitalcorpora.org/, accessed 2022-05-15.
[8] https://www.cfreds.nist.gov/, accessed 2022-05-15.
[9] https://www.iblue.team, accessed 2022-05-15.

Table 3
For each dataset we counted the frequency of accounts (Acc), contacts (Con), messenger apps (Msgr), messages (Msgs), calls, geospatial data (Geo), databases (DB), pictures (Pic), video files (Vid), audio files (Aud) and documents (Doc) according as defined in Section 3.2.

ID	Device	Acc	Con	Msgr	Msgs	Calls	Geo	DB	Pic	Vid	Aud	Doc
1	Samsung Galaxy J3	12	21	2	56	3	10	294	1175	0	3	4
2	Motorola Milestone	0	0	0	0	0	0	1874	215	0	83	132
3	Motorola Droid	5	8	7	18	0	29	822	1063	0	177	872
4	HTC S620	0	0	1	13	0	0	1	387	5	16	53
5	Nexus S	2	20	3	19	8	31	64	289	0	67	459
6	Apple iPhone 3G	0	0	1	6	0	43	113	11796	20	431	1244
7	LG Nexus 5	16	3	7	20	0	0	215	11501	62	230	1391
8	Google Pixel 3	5	26	26	84	11	6	874	21289	208	892	1051
9	LG Nexus 5X	6	11	16	474	8	8	566	8846	91	506	14985
10	LG Nexus 5X	2	12	17	98	14	1	514	25346	80	309	24809
11	Google Pixel 3	4	21	18	89	12	4	377	12339	38	416	772
12	Google Pixel 3	51	57	25	453	68	16999	1256	795	39253	304	18063
13	iPhone SE	18	0	29	218	32	902	891	28931	283	1731	2798
14	iPhone SE	18	0	30	235	37	917	866	27730	298	1838	1834
15	Nexus One	0	0	0	0	0	3	43	121	0	90	13
16	Nexus S	0	0	0	0	0	2	34	107	0	65	13
17	iPhone 3 GS	0	0	2	21	0	11	15	66	0	48	27
18	LG Optimus	0	0	0	0	0	0	1397	16	0	81	1401
19	Samsung Galaxy S4	26	11	10	11	28	3	360	10384	49	293	1355
20	HTC Desire 626s	17	11	7	2	2	2	226	875	17	297	514
21	LG K7	35	17	9	11	0	2	250	6011	4	253	831
22	Motorola Moto-E (2G)	0	0	0	0	0	0	2055	407	24	118	181
23	ZTE ZMax	43	19	9	21	0	0	237	11984	16	269	845
24	HTC Desire S	16	8	5	13	0	11	214	526	19	89	854
25	Samsung Galaxy S2	27	8	6	18	33	2	210	602	8	232	182
26	HTC One XL	14	7	6	15	0	0	203	693	30	218	865
27	HTC One Mini	114	45	9	63	1	0	288	8779	52	236	1408
28	Google Pixel 3	3	2	5	98	0	50	324	4538	104	43	10850
29	Samsung Galaxy Note 10	22	28	11	263	10	103	2012	22175	119	590	5740
30	Apple iPhone X	14	82	8	605	55	28644	928	26171	560	2123	4671
31	Apple iPhone X	0	0	2	0	0	1321	1372	58755	2017	5157	7141

(x64-based, Build, 19470) machine running on Intel Core i5-8265U with 16 GB DDR4 2133 MHz RAM. Autopsy comes with a set of modules specialized for dedicated digital forensic tasks. The overview and the versions of our Autopsy modules are the following: *Recent Activity* (4.19.1), *Virtual Machine Extractor* (4.19.1), *iOS Analyzer (iLEAPP)* (4.19.1), *Android Analyzer* (4.19.1), *File Type Identification* (4.19.1), *Embedded File Extractor* (4.19.1), *Email Parser* (4.19.1), *Extension Mismatch Detector* (4.19.1), *Interesting Files Identifier* (4.19.1), *PhotoRec Carver* (7.0), *Picture Analyzer* (4.19.1), *Central Repository* (4.19.1), *GPX Parser* (1.2) and *Data Source Integrity* (4.19.1).

However, we are not able to extract information from all forensic datasets. We therefore complement our Autopsy-based analysis in four cases, where Autopsy fails to work. This holds for the datasets ID 1 (Samsung Galaxy J3) and IDs 29–31 (Samsung Galaxy Note 10 and two Apple iPhone X), details on the datasets are given in Tables 1 and 3, respectively. These datasets are only available in the UFED format, which is proprietary for Cellebrite products Cellebrite DI Ltd. (2022) and is currently not supported by other forensic tools. Therefore, we analyzed the dataset IDs 1, 29, 30 and 31 with the UFED Reader/Physical Analyzer version 7.44.0.80 and processed the results in the exact same manner.

To maximize the number of data found, we also applied the modules to identify and carve files from unallocated space and compressed archives, i.e. for Autopsy the *File Type Identification*, *Embedded File Extractor* and *PhotoRec Carver* modules. File carving is a common approach to determine the file type, when either the file system is corrupt or missing. By checking for specific byte combinations in header and footer data blocks of files, the data type can be determined, even when the file extension (e.g. *.jpg, *.pdf) is missing or changed (e.g. *.0, *.custom) Lin (2018). This results for Autopsy counting all occurrences of (carved) files inside an image, including non-user generated files, i.e. system and app files, which typically do not change, even when a user interacts with the device.

3.2.2. *Analysis preprocessing*

Each mobile forensic dataset in Table 1 corresponds to a particular device and, if applicable included the corresponding external storage, typically a Secure Disk Card (SD Card). Further, some sources provided multiple extraction methods for one device, stored as separate image files. The image files may yield different content, even when there was no user interaction Ahmed and Dharaskar (2008) or they were created with different methods. In our analysis, we only considered the extraction method with the potentially highest amount of information, i.e. in descending order an extraction based on chip-off, JTAG, physical, file system and logical extraction as described by Alghafli et al. Alghafli et al. (2012) and in the Android training manual of the California Cybersecurity Institute by Elwell and Poirier (2019). In case that multiple versions of the same extraction method are available (, e.g. one web resource provided images on a daily basis with slightly different content), we only examined the latest image file, which potentially holds the most information. Furthermore, we have to discard from our gathered corpora list the mobile datasets within the NIST CFReDS Mobile Archive[10] as this web resource mainly contains information about different forensic tool reports without providing the original contents. However, we managed to extract the contents of the dataset ID 17 (iPhone 3 GS). A final preprocessing point is that we exclusively executed the Android module on Android devices and the iOS Analyzer (iLEAPP) on iOS devices, respectively. As already explained, we decided to use an open-source tool in our analysis, so the forensic community can reproduce our results without the

[10] https://cfreds-archive.nist.gov/mobile/mobile-archived-images.html, accessed 2022-05-15.

need of expensive proprietary software.

3.2.3. Analysis results

In the following we provide our analysis results of publicly available mobile corpora. Our analysis prepares our assessment of the available mobile forensic datasets presented in Section 4.

For our analysis we do not label the content, i.e. we do not distinguish of what could be considered relevant or irrelevant information in a dataset's scenario, respectively. We present the results of the raw output from our analysis environment (i.e. mainly the output of Autopsy and its modules as described above). For each data category we calculate statistical parameters as shown in Table 2. The statistical parameters may be considered as a starting point to create additional datasets. In what follows we discuss the statistical values for the selected data categories.

3.2.4. Final remarks

The iPhone SE datasets run on iOS versions 13.3 and 13.4, respectively. Although, the two datasets describe a similar scenario, we decided to include them both in our analysis, as the second device was used, according to the accompanying documentation for 3 more days. Notable, is the relative high number of geospatial data in IDs 12 (Google Pixel 3), both iPhone SEs (IDs 12 and 13) and both iPhone Xs (IDs 30 and 31), in contrast to the other mobile datasets, which hardly store any geospatial data. The device ID 30 stores more than 28 thousand geographic data points (first most data points of all mobile datasets) followed by ID 12 with nearly 17 thousand from last known locations and extracted from EXIF tags.

In contrast to the datasets 12–14, 29 and 30, we find several mobile datasets with little to no data. This is the case in the *Nexus Experiments*, found on the *Digital Corpora* website, named Nexus One/S (image IDs 15 and 16). The same applies for the device HTC S620 (ID 4), a device that runs on the Windows Phone OS, where the used modules do not support this operating system (OS). Further, we encounter multiple devices, where Autopsy cannot reconstruct single partitions or the file system, because according to their documentation on the Android module, they do not support interpreting older Android file systems Carrier (2022). Therefore, we fail to consider this data. This was the case for images Motorola Milestone (ID 2, DFRWS FC), LG Optimus (ID 18, NIST CFReDS) and Motorola Moto-E 2nd Gen (ID 22, NIST CFReDS). Remarkably, two out of three devices are manufactured from the same company ‘Motorola’, which may point to a misinterpretation of company-specific builds.

4. Assessment of mobile forensic datasets: quantity, quality and timeliness

Section 3 shows the results of our search and of our analysis of publicly available mobile forensic datasets. In this section, we tie in with our analysis and assess these datasets according to quantity, quality and timeliness in contrast to a commonly used smartphone. Our assessment discussion is given in Section 4.2. However, we first aim at introducing our understanding of a ‘typical’ dataset for use in forensics in Section 4.1.

4.1. A realistic mobile dataset

A typical representative of a current mobile device is a smartphone having either Android or iOS as its operating system. This device is used for multiple purposes, in particular integrated in the user's everyday tasks. Therefore, the device contains multiple applications installed to fit the user's needs. Further, the user's interactions leave traces in form of files and as entries in various databases, e.g. stored pictures taken with the camera app. Larger internal and cloud storage abilities encourage the user to store and generate contents at a large scale without needing to explicitly deleting single files. Therefore, devices being used for longer periods tend to use nearly the whole capacity of the storage device. Assuming now, that this user would additionally commit a crime while using the same mobile device, then data in the scope or time frame of the criminal action is stored as a trace on the device, too. On the first sight, a mobile device being a witness in a crime does not differ from a regular device without proper analysis. We use this set-up of a current mobile device in the following to assess the mobile forensic datasets in Section 4.2. Hence, mobile devices exclusively used for criminal purposes (e.g. a smartphone only used for calls with peers) is not addressed by our evaluation.

4.2. Assessing the average dataset

In this section, we assess our gathered datasets according to quantity, quality and timeliness in relation to a commonly used smartphone as introduced in Section 4.1. Our key result is given in Table 1, where we summarize information about the mobile corpora along with the operating system, extraction method, year of creation and a rating in terms of quantity and quality. While the assessment categories quantity and quality are proposed by Grajeda et al. (2017), we extend them by a further category timeliness to ensure that the corpus is not outdated. A published rating of mobile forensic corpora with respect to all evaluation categories does not yet exist to the best of our knowledge. In the following, we choose the mean values per category described in Table 2 and call it the *average dataset* and compare the averages in contrast to a current mobile device as described in the former Section.

4.2.1. Assessing the mean values

We observe 15 *accounts* in average, which could also be found in a current mobile device as it considers user credentials of social media, online accounts and messenger apps. Assuming, that from 15 accounts 9 are from *messenger apps* and therefore leaving with 6 accounts from other services. According to an online survey about smartphone usage from GlobalWebIndex Mander and Kavanagh (2019) the average smartphone user has 6.4 social media accounts and hence, this number seems to be realistic to us.

The average dataset has 13 *contacts* stored on a device. This number is very low, because a typical user may store contact information including the phone number, email addresses and mailing addresses on his regular social contacts. Therefore, we can assume a much higher number in a realistic mobile device.

Likewise, having 10 calls on average is also rather low, considering that this number represents all incoming and outgoing calls over a large period, even though it does not count voice calls over messenger apps.

Also, in total 94 *messages* over the cellular network and through messenger apps is again very low in quantity. Current mobile devices may contain 10000s or more messages using modern messenger apps.

The average dataset contains 1584 entries with *geographical data points* (geodata), although modern smartphones usually track their current location and is then used to optimize various online services. To name one service: semantic search queries used by popular search engines often include the current location in their search to optimize the search results. In addition, the camera apps often store geodata inside EXIF tags in the picture's file header, too. Hence, we assess to find too few geodata traces in the average dataset and find only two datasets (IDs 12 and 30) having more geodata than the average.

In a similar manner, we account for 1399 *video files* in the average dataset and again only two datasets (IDs 12 and 31) having more video files than the average dataset. These datasets might better resemble current mobile devices, because modern camera apps also store a short footage of the picture's time context.

Moving on, there is a high mean number of *photos* and *pictures* with 9804 files, even when considering that most online applications and websites may cache some pictures on the device's memory to minimize the amount of data transferred over the cellular network. The average dataset thus contains a sufficient number of pictures.

Having 3399 user-generated *documents* on a mobile device is very high, but considering, that the number includes all files with any kind textual information then we can reconsider it realistic. The quantity of 610 *databases* seems to be a sufficient quantity, too. When considering that on average 555 *audio files* are stored on the device we can derive this number being realistic as well.

To sum up, most categories do not suffice to be considered rich in quantity and therefore not be considered a good representative for a current mobile device. This underlines the poor quality of available mobile forensic corpora.

4.3. Assessing each dataset

While the average dataset can be used as reference in comparing different datasets, it does not give an objective means to choose a single dataset for training and validation of forensic tools. Therefore, we rate each dataset according to a definition of our assessment categories *quantity*, *timeliness* and *quality*. The result is given under the columns *quantity*, *year* and *quality* in Table 1.

4.3.1. Quantity

The huge variance in the datasets is an indication towards different experimental set-ups and goals in the creation of the data. In order to rate the quantity for a mobile corpora, we calculate the lower bound of the upper quartile (i.e. the 75th percentile) for each category from the raw data in Table 3. This threshold states that three quarters of the data points are below this value and therefore a quarter of the data points above. In our case with 31 data points, we can then identify the 8-highest values, if these are higher than this threshold.

Exemplified for the *geodata* category: the 75th percentile for geodata is 1584, hence the 8 datasets with IDs 6, 12–14 and 28–31 contain more than 1584 geodata.

We then count for each mobile dataset how many categories contain more data points than the threshold for the respective category. With this approach, we can simply categorize datasets by their quantity. We make use of a three class rating to identify three types of datasets: datasets with overall little content (0–3 categories with the 8th most data points), datasets with many contents for some categories (4–7 categories with 8th most data points) and those with overall many contents (8–11 categories with 8th most data points). We label these datasets with the corresponding symbol: “-” for poor, “+/−” for mixed and “+” for datasets with rich contents in Table 1.

Only the most recent datasets 12 (Google Pixel 3), 13/14 (iPhone SE), 29 (Samsung Galaxy Note 10) and 30 (iPhone X), can be considered datasets with a decent amount of content. The datasets 8 (Google Pixel 3), 10 (LG Nexus 5X) and 31 (Apple iPhone X) have rich contents for some categories and may point to biased data, e.g. the dataset 31 contains very high number of picture and audio files, but appears not having any cellular activity.

Surprisingly, the Forensic Challenges (IDs 2–7) do not contain sufficient data on a single device having contents rich in quantity. A reason for this may be that the scenarios include various devices in their experiments and therefore the information is split across multiple devices. In addition, we identify that the *Digital Corpora* and *CTF scenarios* contain in general richer content than the other sources. To sum up, most of the publicly available mobile forensic datasets contain too few traces to be considered as realistic.

4.3.2. Timeliness

The assessment category *timeliness* is a unit, which denotes the age of a dataset. A recently created dataset is a good representative of data found in current devices, as contemporary applications and devices use current technologies and apps to store their data. We extract from Table 1 that out of the total 31 mobile datasets only 22 were created within the last five years. Furthermore, only 13 of the mobile datasets are not older than three years. Hence, not even half of the published corpora actually represents a current device. To sum up, timeliness is not ensured by currently available published mobile forensic datasets.

4.3.3. Quality

To asses the quality of a dataset, we combined the values of quantity and timeliness and rated the dataset respectively in the same manner as for quantity: in good, neutral and poor quality datasets. We define a high quantity rating as follows: the combination of a high timeliness and a high *quality* rating, i.e. a dataset created in the last 5 years, which contains a high quantity of data leads to a high quality rating. Respectively, having a low quantity rating and the dataset being older than 5 years results in a low quality rating. Finally, a recent dataset with few traces is rated as neutral. An exception to this rule is, when we could not reconstruct the files or the file system due to corrupted images or files. In this case, we gave the dataset a poor quality rating. In detail, this was the case for dataset IDs 2–7, 15–18 and 22. In all, we only have 5 corpora of high quality, from which two of them provided by the Cellebrite's CTF event Cellebrite DI Ltd. (2022). In our opinion this proves the mobile forensic dataset problem.

5. Perspective of mobile forensic corpora

In Section 4 we have shown the existence of the mobile forensic dataset gap. In the following, we shortly discuss how to proceed to fill this gap.

5.1. Real datasets vs. synthetic

In Section 2 we reviewed the generation of (mobile) forensic corpora. As of today researchers can use either data manually created from experiments (real corpora) or from computer simulations (synthetic corpora).

Real corpora might generate more realistic data, because it is generated by real user interactions. This approach can quickly generate small amounts of data, without needing any specialized software. In general, any user with a basic knowledge how to operate mobile devices might easily generate a vast number of traces. However, the use and generation of real corpora may be limited by physical constraints (, e.g. collecting data of restricted areas), resource limitations (, e.g. limited in time or funding) or legal restrictions (, e.g. personal data are restricted by privacy laws such as the European Data Regulation by the European Parliament and Council (2016)).

Synthetic corpora, on the other hand, are generated by running computer simulations to leave traces of simulated user interactions on a device. This is achieved by either manipulating files or placing bits of evidence on the device's storage. Synthetic corpora can be easily scaled to meet the desired quantity with an ordinary computer. In addition, a synthetic approach excludes that the results do not contain any personal data and therefore enabling researchers to freely share their data with others.

Another persistent problem in the forensic community is the missing of a ground-truth and labeled data in real datasets. On the other hand, a synthetic dataset may be generated using a set of configuration files, that contain a description of the scenario, i.e. the ground-truth. As a result, the validation of a forensic tool (, e.g. as proposed by National Institute of Standards and Technology) may be done with a synthetic dataset, because a ground-truth is given.

One obvious approach to solve the missing dataset problem, is to focus future research in extending existing forensic data synthesis tools like *TraceGen* Du et al. (2021), *hystck* Göbel et al. (2020), *FADE* Delgado et al. (2021) or ForTrace Göbel et al. (2022) to include the support of popular mobile devices, in particular devices running Android and iOS.

An important issue in this context is the transfer from the mobile testing community to the digital forensic community. For instance the Python-based Android View Client[11] seems a comfortable emulator from the Android testing community to generate forensically relevant traces in an automatic way.

5.2. Generation of forensic content

In addition, the contents of the dataset may also be synthetically generated, in particular when planning to share the dataset with the community. The generation of synthetic contents, e.g. chat messages, video files, pictures or geospatial data are non-trivial and still part of current research. The labeling process, e.g. in task-relevant and irrelevant information, is highly dependent on its context and should therefore be integrated as preset information in the generation process.

5.3. Requirements

We conclude our perspective section by pointing to requirements on an image generator to actually solve the dataset problem. Besides the requirements already stated in the work by Grajeda et al. (2017) (i.e., *availability*, *quantity* and *quality*) we support the importance of the following requirements on a dataset/forensic image generator as described by Göbel et al. (2022):

- *Timeliness*: The dataset contains traces embedded in a contemporary environment. Having outdated datasets does not give realistic insights in current problems due to the short lifetime and update cycles of hardware devices, operating systems and used applications.
- *Adaptability*: A forensic image generator must produce customizable datasets to meet individual specifications and thus easing the generation of similar datasets.
- *Labels*: Knowing a ground-truth is crucial for educational purposes and forensic tool testing. Labeled content provide additional information about the dataset's contents and it's environment.
- *Indistinguishability*: Any synthetic generated data-set must represent contents similar to those of contents found in real devices. The forensic expert or tool inspecting the synthetic dataset must not be able to distinguish it from a real dataset.

Furthermore the *holistic* aspect as explained by Göbel et al. (2022) should be considered as a key requirement of a mobile forensic data set, too. Especially the generation of a corresponding network capture is not addressed yet.

6. Conclusion and future work

We reviewed for publicly available mobile forensic corpora and created an overview where to find them and additionally summarized for each dataset its contents with respect to 11 categories. Further, we rated each dataset for quantity, quality and timeliness. Most of the available mobile corpora do not suffice to be considered rich in quantity and mostly contain system and app data rather than user-generated content. Only 13 out of 31 datasets were generated in the past three years and older datasets may contain deprecated versions of applications and operating system builds. Therefore, most of the mobile corpora do not contain high-qualitative data. We hence showed that the mobile dataset gap actually exists. Thus, the lack of having sufficient available mobile forensic corpora is still an issue.

Furthermore, the research focus of the forensic community should not only include the creation of more mobile forensic corpora, but also suffice in quantity, quality and timeliness. Rather than just appealing to researchers in sharing their manually generated real dataset with the community, we further suggest to focus the research on creating tools and frameworks for generating synthetic mobile forensic corpora. These synthetic corpora easily scale in quantity in contrast to real datasets created from experiments and follow additional requirements of timeliness, adaptability, indistinguishability and existing ground-truth (labels).

Future research should hence extend or develop forensic image generators to support the generation of synthetic mobile corpora and also synthetic forensic content. An important step is the review of mobile testing frameworks like Android View Client and their capabilities to contribute to a synthetic digital forensic image generation framework. The actual content should be realistic, labeled, up-to-date and not containing personal data. The results should break the circle of constantly lacking corpora for validating tools and training experts on the field.

References

Abt, S., Baier, H., 2014. Are we missing labels? A study of the availability of ground-truth in network security research. In: 2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS). IEEE, pp. 40–55.

Ahmed, R., Dharaskar, R.V., 2008. Mobile forensics: an overview, tools, future trends and challenges from law enforcement perspective. In: 6th International Conference on E-Governance, ICEG, Emerging Technologies in E-Government. M-Government, pp. 312–323.

Alghafli, K.A., Jones, A., Martin, T.A., 2012. Forensics data acquisition methods for mobile phones. In: 2012 International Conference for Internet Technology and Secured Transactions, pp. 265–269.

Beebe, N., 2009. Digital forensic research: the good, the bad and the unaddressed. In: IFIP International Conference on Digital Forensics. Springer, pp. 17–36.

Carrier, B., 2005. Digital forensics tool testing images. http://dftt.sourceforge.net/. (Accessed 11 May 2022).

Carrier, B., 2022. Autopsy - the sleuth kit. https://www.sleuthkit.org/. (Accessed 11 May 2022).

Cellebrite DI Ltd, 2022. Cellebrite UFED analyzer. https://www.cellebrite.com/. (Accessed 11 May 2022).

Delgado, A.A.C., Glisson, W.B., Grispos, G., Choo, K.K.R., 2021. Fade : a forensic image generator for Android device education. WIREs Forensic Science. https://doi.org/10.1002/wfs2.1432.

Digital Corpora, 2022. http://digitalcorpora.org/corpora. (Accessed 11 May 2022).

Du, X., Hargreaves, C., Sheppard, J., Scanlon, M., 2021. TraceGen: user activity emulation for digital forensic test image generation. Forensic Sci. Int.: Digit. Invest. 38, 1–11. https://doi.org/10.1016/j.fsidi.2021.301133, 301133.

Elwell, C., Poirier, J., 2019. Windows and Android Forensics Ccic Training. https://cci.calpoly.edu/2019-digital-forensics-downloads. (Accessed 11 May 2022).

European Parliament and Council, 2016. Regulation (eu) 2016/679 of 27 April 2016. Off. J. Eur. Union 59, 1–88. http://data.europa.eu/eli/reg/2016/679/oj.

Garfinkel, S., 2007. Forensic corpora: a challenge for forensic research. Electron. Evid. Inf. Cent. 1e10.

Garfinkel, S., Farrell, P., Roussev, V., Dinolt, G., 2009. Bringing science to digital forensics with standardized forensic corpora. Digit. Invest. 6, S2–S11.

Garfinkel, S.L., 2010. Digital forensics research: the next 10 years. Proc. Tenth Annual DFRWS Conf. 7, S64–S73. https://doi.org/10.1016/j.diin.2010.05.009. the

[11] https://pypi.org/project/androidviewclient/, accessed 2022-05-15.

Proceedings of the Tenth Annual DFRWS Conference. https://www.sciencedirect.com/science/article/pii/S1742287610000368.

Göbel, T., Maltan, S., Türr, J., Baier, H., Mann, F., 2022. ForTrace - a holistic forensic data set synthesis framework. In: DFRWS EU 2022, 301344.

Göbel, T., Schäfer, T., Hachenberger, J., Türr, J., Baier, H., 2020. A novel approach for generating synthetic datasets for digital forensics. In: IFIP International Conference on Digital Forensics. Springer, pp. 73–93.

Grajeda, C., Breitinger, F., Baggili, I., 2017. Availability of datasets for digital forensics–and what is missing. In: Proceedings of the Seventeenth Annual DFRWS USA, 22, pp. S94–S105 (S).

Lillis, D., Becker, B., O'Sullivan, T., Scanlon, M., 2016. Current challenges and future research areas for digital forensic investigation. In: Eleventh Annual ADFSL Conference on Digital Forensics. Security and Law, pp. 9–20.

Lin, X., 2018. File carving. In: Introductory Computer Forensics. Springer International Publishing, Cham, Switzerland, pp. 211–233.

Luciano, L., Baggili, I., Topor, M., Casey, P., Breitinger, F., 2018. Digital forensics in the next five years, in: proceedings of the 13th international conference on availability, reliability and security. New York, NY, USA Assoc. Comput. Mach. 1–4. https://doi.org/10.1145/3230833.3232813.

Mander, J., Kavanagh, D., 2019. Social media flagship report: online research among Internet users aged 16-64. Research report. GlobalWebIndex. https://www.gwi.com/hubfs/Social%20Report.pdf.

Moch, C., Freiling, F.C., 2009. The forensic image generator generator (forensig2). In: 2009 Fifth International Conference on IT Security Incident Management and IT Forensics, pp. 78–93.

National Institute of Standards and Technology. Computer forensics tool testing program. https://www.cftt.nist.gov. (Accessed 27 January 2022).

Scanlon, M., Du, X., Lillis, D., 2017. Eviplant: an efficient digital forensic challenge creation, manipulation, and distribution solution. Digit. Invest. 21.

Trail of Bits, 2022. Forensics. https://trailofbits.github.io/ctf/forensics/. (Accessed 11 May 2022).

Visti, H., Tohill, S., Douglas, P., 2015. Automatic creation of computer forensic test images. In: Computational Forensics. Springer, pp. 163–175.

Woods, K., Lee, C.A., Garfinkel, S., Dittrich, D., Russell, A., Kearton, K., 2011. Creating realistic corpora for security and forensic education. In: Sixth Annual Conference on Digital Forensics. Security and Law, Monterey, CA, USA, pp. 123–134.