Chapter 3

SMARTPHONE DATA DISTRIBUTIONS AND REQUIREMENTS FOR REALISTIC MOBILE DEVICE FORENSIC CORPORA

Patrik Goncalves, Andreas Attenberger and Harald Baier

Abstract Mobile devices such as smartphones are carried and used constantly by people in their daily lives and, therefore, play important roles in forensic investigations. As a result, digital forensic professionals are confronted with large numbers of devices with data that has to be extracted and analyzed. The education and training of forensic experts and the development and evaluation of smartphone forensic tools require copious amounts of realistic data. Unfortunately, secrecy and privacy considerations limit the availability of real digital forensic data. Smartphone datasets for training and testing are sparse and unrealistic, and knowledge about data distributions in real smartphones is limited.

This chapter presents the results of a survey of law enforcement professionals from two countries that sought to understand the typical data residing in smartphones encountered in criminal investigations, with the goal of supporting the creation of publicly-available forensic datasets. The typical data extracted from smartphones using current forensic tools is presented; this data is divided into two forensic classes, relevant and irrelevant. Additionally, the chapter discusses current problems encountered by mobile device forensic professionals and opportunities for future research.

Keywords: Smartphones, investigations, datasets, data distributions

1. Introduction

Mobile devices such as smartphones, tablets, wearable computers and Internet of Things devices are carried and used by people in their daily lives and, therefore, play important roles in forensic investigations. In fact, law enforcement professionals encounter increasing numbers of portable devices compared with stationary devices such as desktop comput-

Published by Springer Nature Switzerland AG 2022
G. Peterson and S. Shenoi (Eds.): DigitalForensics 2022, IFIP AICT 653, pp. 47–63, 2022.
https://doi.org/10.1007/978-3-031-10078-9_3

ers in forensic investigations [2]. Since smartphones incorporate embedded sensors and enable users to adapt their functionality by installing custom applications, they contain valuable information about user activities and their contexts (e.g., business, social and criminal contexts) [13].

Law enforcement professionals employ a number of forensic tools to extract and analyze relevant information from smartphones. Since the results are included in final reports presented in court proceedings, it is mandatory that the forensic tools are validated against realistic test datasets to minimize false positives and false negatives. Unfortunately, using real datasets drawn from confiscated devices to validate forensic tools is not an option due to ongoing investigations [15] and data protection regulations such as the European Union's General Data Protection Regulation [5]. Some researchers have clearances or authorizations that enable them to use real data, but this is a very small group and the results may not be available for public use. As a result, the only options available to the digital forensics community are to use data from public sources or create their own training and testing datasets.

Synthetic forensic data, like real forensic data encountered in investigations, should convey scenarios involving criminal and non-criminal activities. Defining realistic scenarios is not easy because it requires detailed knowledge about criminal and non-criminal behavior and how smartphones are used in these contexts. Realistic scenarios are best created by interacting with experts, especially law enforcement professionals with extensive experience extracting and analyzing data from seized smartphones.

Complex forensic scenarios are typically created and published by dedicated working groups that draw on the knowledge and experience of experts. A key drawback of published datasets is that their scenarios and contents do not change. Researchers with resources and time often create their own forensic datasets to suit their needs. However, Grajeda et al. [8] report that such datasets are shared in a limited manner or not shared at all.

A digital forensic professional uses various forensic tools to obtain all the information that is available to answer a set of investigative questions. Some of the information is labeled as relevant and the rest is labeled as irrelevant. The labeling process is not simple; it depends greatly on the specific investigation and the information recovered in the investigation. Thus, a researcher who seeks to create a synthetic forensic dataset must know what constitutes typical relevant and irrelevant information and where the information resides in smartphones.

An important use case for a forensic dataset is to validate forensic tools against the ground truth. The ground truth is expressed by correctly-

labeled data corresponding to the categorization of digital artifacts into task-relevant and task-irrelevant data. Unfortunately, Abt and Baier [1] note that publicly-available datasets often do not provide ground truth data. This missing labeled data problem hinders research and development efforts focused on novel forensic tools and methods. The absence of labeled data also hinders the comparability and repeatability of results.

Technological advances make it particularly challenging to keep up with the data content of smartphones. Smartphone content can be categorized by file class (picture, video or document file) or entry class found in one or more files (contacts, chat messages, browser log entries or geospatial data). Having collected all the content in a smartphone, it is possible to state that the device contains certain numbers of files and entries with certain statistical distributions. Abt and Baier [1] note that statistical properties may be used to assess the quality of synthetic data sets with respect to real data. The statistical properties may also be extended to assess manually-created data sets.

Employing a statistical approach requires knowledge about the data distributions in smartphones, but this knowledge is currently missing. In order to address the problem, this research focuses on acquiring knowledge about smartphone data distributions. Specifically, a survey was conducted of law enforcement professionals to obtain detailed information about smartphone data content. All the survey participants were digital forensics experts who actively worked on extracting and analyzing smartphone data.

The survey focused on the contents of a representative smartphone that would be acquired and analyzed by a law enforcement professional to gain insights into data statistics. The representative contents were labeled into typical task-relevant and task-irrelevant content. The survey also attempted to understand the problems encountered by the participants while conducting forensic examinations of smartphones with the goal of articulating law enforcement needs related to mobile device forensics.

The research results are intended to assist the mobile device forensics community in creating synthetic datasets and employing statistical properties to compare the synthetic datasets against the data contained in real mobile devices. The insights gained into the problems encountered by law enforcement professionals are intended to enable the mobile device forensics community to help address current and future law enforcement needs.

2. Related Work

This section discusses work published between 2010 and 2021 on assessing the problems encountered by digital forensic professionals and their needs related to mobile device forensics. The section also discusses efforts focused on creating realistic mobile device datasets for forensic tool testing and evaluation. The literature review leveraged four leading research publisher databases and two scientific research search engines:

- **IEEE Xplore** (`ieeexplore.ieee.org`).
- **ACM Digital Library** (`dl.acm.org`).
- **Elsevier ScienceDirect** (`www.sciencedirect.com`).
- **Springer Link** (`link.springer.com`).
- **Google Scholar** (`scholar.google.com`).
- **ResearchGate** (`www.researchgate.net`).

In 2010, Garfinkel [6] published his seminal work on the state of the art of digital forensics and the future of digital forensics research. He described the challenges that existed and how to make future research in digital forensics more efficient. A key finding was the absence of standardized file formats and forensic tools, which needed to be addressed by collaborative efforts involving practitioners, researchers and industry. However, while this work mentioned mobile device forensics, it did not address the need to understand and process large amounts of diverse mobile device data.

Motivated by the challenges faced by network forensic practitioners, Woods et al. [15], in 2011, published their experience creating the M57-Patents dataset comprising realistic traffic involving multiple networked devices. The M57-Patents dataset was the result of a workshop attended by several experts who created a realistic scenario with criminal activity. The primary objectives were to provide answer keys (ground truths of scenarios) with realistic digital artifacts generated from applications, networking and background processes. The resulting disk images, traffic dumps, RAM dumps and other evidentiary data were published to advance network forensics education and training. The work of Woods and colleagues has motivated this research focused on smartphone data content and distributions.

In 2016, Lillis et al. [9] identified technical challenges in the digital forensics domain based on an extensive review of the contemporary literature. A key problem for digital forensic professionals was coping with

the numbers, heterogeneities and data volumes of mobile devices. They also discussed the need to address the interactions of mobile devices with other data sources such as Internet of Things devices and cloud resources. However, Lillis and colleagues did not consider smartphone content and the statistical distributions of smartphone data.

In 2018, Luciano et al. [11] described the results of a workshop attended by digital forensic professionals that sought to identify important research issues over the next five years. Their findings included limited research funding, absence of standards, limited multidisciplinary knowledge and approaches, lack of information sharing and collaborative activities, use of outdated techniques and tools, and the need to advance the reputation of digital forensics as a discipline.

In 2018, Barmpatsalou et al. [2] published a review of the contemporary literature on mobile device forensics. They noted that data encryption was more prevalent and that it was much harder to decrypt data. Also, the diversity of devices, operating systems and software was creating incompatibilities with commercial forensic tools, requiring digital forensic professionals to pursue manual efforts or use third-party software. Additionally, they encouraged tool developers to focus on the standardization of forensic file formats and tool interoperability. A key critique was that researchers were not focusing on automated methods for evidence classification. The gaps included data and artifact classification, user behavior pattern detection, automated malicious activity detection, multi-source data correlation and criminal activity detection by analyzing data patterns.

Also in 2018, Camacho et al. [3] published a review of contemporary mobile device forensics. They identified the lack of standardized methodologies and the need to use large numbers of forensic tools to achieve investigative goals. Additionally, they noted the need to integrate browser applications that support instant messaging, social networking, email, video and audio analysis.

3. Survey Methodology

This research sought to capture the knowledge and experience of law enforcement professionals related to the forensic extraction and analysis of data from seized smartphones and specify the contents of a representative smartphone encountered in criminal investigations.

Goals. A representative smartphone would provide digital forensic researchers with valuable insights pertaining to the forensic analysis of devices seized in criminal investigations. These include the type and amount of data at the file-class level (e.g., databases, pictures, videos,

audio and text documents) as well as the data structures encountered in multiple file classes (e.g., accounts, contacts, messenger apps, calls and geospatial data). The survey focused not on the actual file extensions (e.g., PNG, JPG and GIF in the case of pictures), but the types of content that are represented (e.g., when file carving is used to determine file types [10]).

To cope with the data labeling problem [1], the survey participants were asked to classify content into two classes, typical task-relevant information and typical task-irrelevant information. The survey participants were also interviewed to identify the problems encountered while conducting mobile device forensic tasks and law enforcement needs. In summary, the three goals of the survey were:

- **Goal 1:** Gain insights into the data distributions in a representative contemporary smartphone.
- **Goal 2:** Determine notable files and apps that contain large amounts of task-relevant and/or task-irrelevant information.
- **Goal 3:** Identify the problems encountered by law enforcement professionals while conducting mobile device forensic tasks and their needs related to mobile device forensics.

Achieving the first and second goals would assist the digital forensics community in creating realistic smartphone datasets for training and tool testing. Additionally, the representative content provided by the survey participants would support statistical analyses of smartphone datasets and comparisons of data from real devices against synthetic or manually-generated datasets. Achieving the third goal would provide insights into current challenges related to mobile device forensics and steer mobile device forensics research and development efforts.

Survey Design. Interviews were chosen as a qualitative method to assess the knowledge and experience of experts in mobile device forensics [12]. The interviewees were provided with a set of potential questions in advance of the interviews to give them time to prepare their responses. The semi-structured interviews were designed to emphasize extensive discussions and reduce the need for a large survey population. This was deemed necessary because it is difficult to recruit experts in mobile device forensics for research studies due to their high workloads.

Survey Method. Active professionals from different law enforcement agencies whose tasks involved extracting and analyzing smartphone data were selected as survey participants. The interviews were conducted over

an online videoconferencing system to facilitate access. Each interviewee provided a list of crimes that were typically encountered. The interviews were terminated after six participants because thematic saturation of the types of crimes covered was attained.

Survey Questions. The survey had four parts: (i) introduction, (ii) assessing the contents of typical seized smartphones, (iii) assessing the data distributions in a representative smartphone and (iv) open discussion.

During the introduction part, it was ascertained that the participant matched the desired focus group based on occupation, affiliation and personal experience working on smartphone forensics. Also, the forensic tools that were typically used by the participant were identified.

The second part of the survey covered the typical contents of a seized smartphone. Specifically, each participant was asked to identify the data and/or file extensions encountered in large amounts of task-relevant and task-irrelevant content.

The third part of the survey acquired statistical information about the distribution of data in a representative smartphone with respect to 11 categories: (i) account entries (Acc), (ii) contact entries (Con), (iii) messenger apps (Msgr), (iv) text/email messages (Msg), (v) calls made (Call), (vi) geospatial data entries (Geo), (vii) database files (DB), (viii) picture files (Pic), (ix) video files (Vid), (x) audio files (Aud) and (xi) document files (Doc).

The final part of the survey involved an open discussion of topics in mobile device forensics. This included general comments, problems encountered and potential solutions related to mobile device forensics.

Data Collection. During the interviews, all the comments and answers provided by the participants were transcribed directly. The list of questions provided to the participants in advance guided the interviews and facilitated the collection of detailed data. The list also maximized the amount of data collected during the interviews.

Survey Limitations. The interviews were restricted to law enforcement professionals from regional and national agencies in Germany and Switzerland. Surveys of law enforcement professionals from other countries would have to be conducted in the future for validation and generalization. Additionally, the survey responses related to the problems encountered by law enforcement professionals are expected to have a country bias.

Table 1. Statistics of data in a representative smartphone.

	Acc	Con	Msgr	Msg	Call	Geo	DB	Pic	Vid	Aud	Doc
Mean	28	1,161	6	34,759	829	10,232	1,713	182,000	1,571	3,613	8,983
Median	24	270	7	32,253	180	535	825	77,500	788	2,545	8,600
Minimum	10	44	1	10	10	200	200	10,000	50	10	6,000
Maximum	60	4,300	7	150,000	8,000	100,000	12,000	1,300,000	5,000	22,000	15,000

Data Analysis. Responses involving numerical values, such as the data distributions in a typical seized smartphone, were specified as ranges instead of exact values; the means of these ranges were used to simplify subsequent computations. The computed statistical parameters included the mean, median, minimum value and maximum value. The non-numerical responses and comments were accumulated across all the interviews and analyzed in a qualitative manner.

4. Survey Results

Interviews were conducted with six German and Swiss law enforcement officers from different regional/national agencies to obtain a diverse coverage of criminal activities. The crimes investigated by the survey participants included drug crimes, theft, Internet fraud, illegal immigration, property damage, crimes against the state/public, terrorism, weapons, explosives, organized crime, internal affairs, money laundering, murder, homicide and others. This section describes the survey results and compares the collected data against published smartphone datasets.

4.1 Typical Smartphone Content

During the survey, each participant provided data size ratings for 11 data categories: (i) account entries (Acc), (ii) contact entries (Con), (iii) messenger apps (Msgr), (iv) text/email messages (Msg), (v) calls made (Call), (vi) geospatial data entries (Geo), (vii) database files (DB), (viii) picture files (Pic), (ix) video files (Vid), (x) audio files (Aud) and (xi) document files (Doc). The individual ratings were used to compute the mean, median, minimum value and maximum value of each data category. When participants provided intervals instead of single values, the midpoints of the intervals were employed to compute the statistics. The exceptions were computing the minimum and maximum values as the low and high points of the intervals, respectively.

Table 1 summarizes the data distributions in a representative seized smartphone. For instance, the smartphone contains 28 accounts on av-

erage. The median is 24 accounts and the minimum and maximum numbers of accounts are 10 and 60, respectively.

The survey participants emphasized that the data distributions are highly dependent on the specific smartphones. Some smartphones contain little or no data, especially those used exclusively for criminal activities, which might only contain a few text messages. Another factor is the type of crime tied to a seized smartphone. For example, smartphones used for human trafficking, Internet fraud and document forgery tend to have large numbers of stored contacts and messages.

Geospatial data deserves special mention. According to the survey participants, locally-stored geospatial data is very useful in investigations, but popular cloud services such as Google Cloud also store valuable geospatial data. In fact, online services store about 80 times more geospatial data than is stored in a typical smartphone.

4.2 Labeling Typical Content

The survey participants noted that labeling data content as task-relevant and task-irrelevant is not trivial and is highly dependent on the specific case. For example, if the location of a crime is not relevant in a case, then geospatial data is labeled as task-irrelevant. However, in the vast majority of cases, geospatial data is task-relevant.

The survey participants were asked about the data categories that usually contain task-relevant information and those that mostly contain task-irrelevant information. Specifically, the participants had to identify the typical apps and locations of relevant and irrelevant information, respectively. Table 2 summarizes the relevant and exclusively irrelevant content in a representative smartphone as provided by the survey participants.

It is important to note that relevant and irrelevant content are not mutually exclusive; instead, relevant information is a proper subset of irrelevant information. Thus, if a participant deems some type of content to be relevant, then there is a high probability that the content belongs to one of the listed types. On the other hand, if a participant deems some content to be exclusively irrelevant (i.e., absolute complement of relevant information), then the content likely belongs to an exclusively irrelevant type.

The survey participants stated that irrelevant information is often in system and app files; this is typically content that does not change. Furthermore, relevant information is rarely found in general apps used for recreation (e.g., gaming). According to the participants, pictures and videos may be labeled as relevant and irrelevant. Subsets of pictures

Table 2. Relevant and exclusively irrelevant content in a representative smartphone.

	Relevant	**Exclusively Irrelevant**
Data Files/ Structures	Text/email messages, Chats, Calls, Contacts, Geospatial data, Pictures, Videos, Audio (voice messages), Databases	Private/erotic pictures/videos, System/app databases
Apps/ System	Messenger apps, Browser logs, Social media, App usage logs, Wi-Fi logs, Power logs Search queries, Health/fitness apps, Personal notes	Gaming apps, Cookies, System data, Template files
Additional File Types	HEIC (iOS), SQL (DB, SQLITE3), PLIST, PDF, DOC, Arbitrary types (0, DATA, ...)	None

and videos used exclusively for recreation are mainly shared with social network contacts. These files may have erotic (excluding illegal pornography), humorous or informative content and are mostly task-irrelevant.

In contrast, media files (pictures, videos, audio, documents), app databases and content created by user interactions (calls, stored contacts, messages, app usage, geospatial data, notes, search queries and others) often contain task-relevant information. The survey participants stated that geospatial data plays an important role in forensic investigations. They provided examples where analyzing logs containing geospatial data entries provided conclusive evidence in investigations. This information was often found in fitness and health applications or at cloud service providers that tracked and stored smartphone locations and movements.

4.3 Mobile Device Problems and Needs

The survey also focused on the problems faced by the participants while conducting their forensic tasks and solicited information about their needs related to mobile device forensics. The principal findings relate to content extraction, forensic tools and content analysis.

Content Extraction. According to the survey participants, content extraction from smartphones and other mobile devices is often hindered by data encryption or password/PIN locks. A typical example is a seized smartphone whose screen is turned off and a password or PIN is required to unlock the device.

Another problem is that some forensic tools require root access, but this is not always possible due to operating system security mechanisms. Social media applications and cloud services store user credentials on local device storage, but encryption prevents access to the credentials.

Forensic Tools. At this time, no single forensic tool can extract and analyze content in all types of smartphones. The diversity of hardware and operating systems forces digital forensic professionals to use forensic tools from different vendors, each tool with its own proprietary file formats. The smartphone forensic tools used by the survey participants come from Cellebrite (Physical Analyzer, Reader, Pathfinder, UFED, etc.), MSAB (XRY and XAMN), Oxygen Forensics (Oxygen Forensic Detective), Magnet Forensics (AXIOM), X-Ways (X-Ways Forensics), Grayshift (GrayKey) and SQLite Consortium (SQLite). Additionally, the survey participants employ self-developed hardware and software tools for extracting and analyzing smartphone content.

The survey participants noted that tool diversity results in forensic reports being produced in different formats with limited interoperability with other tools. Additionally, the participants, regardless of their affiliations, complained that forensic software, even software procured from leading vendors, tends to have bugs and critical security problems for which patches are rarely provided. This is a concern because the quality and validity of forensic reports could be questioned in court.

The needs of the survey participants include high-quality forensic tools that provide better filtering mechanisms, enable the discovery of correlations in data and support the verification of results (e.g., providing qualitative and/or quantitative evaluations via enhanced user interfaces).

Content Analysis. The survey participants, without exception, stated that they encounter increasing numbers of devices with large volumes of data that have to be extracted and analyzed. Problems are also posed by devices created for use in foreign countries for which content in various languages had to be translated manually prior to analysis. This leads to additional costs for translation as well as delays because survey participants have to wait for translations before they can determine the relevance of content to their investigations.

A related problem involved apps that are commonly used in foreign countries as well as by expatriates in other countries. In many instances, the survey participants were unaware of the app functionality and the information the apps might hold. Even worse, the foreign apps often are not supported by common forensic tools. An example is WeChat, a multi-purpose app from Tencent, that is widely for social networking, instant messaging and mobile payments in East Asia.

Another problem is that common forensic tools support popular apps (mainly communications apps), but may not support other app types such as online booking, shopping and package tracking apps; thus, valuable information in the unsupported apps is not automatically integrated in the final reports produced by forensic tools. Yet another problem encountered by the survey participants is recovering and recreating information from deleted encrypted files and deleted SQLite database entries.

Finally, the survey participants observed that criminals deliberately inject files with false content in their smartphones. The anti-forensic files, which come from various sources, mimic content created by the smartphones, deceiving forensic professionals to classify the files as false positives and covering criminal activity. One example given by a survey participant was the injection of pictures that apparently show illegal weapons. The pictures of the weapons, obtained from public sources, were recreated on the smartphone using its camera app to show that the suspect was at the locations where the pictures were originally taken.

5. Discussion

This section discusses the principal findings of the survey. Also, it compares the statistical properties of the data categories in the representative smartphone created by the survey against the statistical properties of the data categories in a published dataset.

5.1 Key Findings

The assessment of the problems and needs with regard to mobile device forensics pointed out the deficiencies in current forensic tools. The problems mentioned by the survey participants were similar to those identified by the digital forensics community in the past. Problems related to forensic tools that persist are the lack of standardized forensic file formats and interoperability between forensic tools [2, 6, 9, 11], bugs in forensic tools and long update cycles [6, 11] and lack of tool support for efficient identification of relevant information (e.g., data correlation and pattern detection) [6, 9].

Validation and verification of forensic tool results based on qualitative and/or quantitative feedback are also problems that persist [6]. Current commercial tools are still black boxes without any documentation about their internal algorithms. Absent adequate validation and verification, forensic professionals and courts are forced to blindly trust commercial tools although validation and verification are important components of forensic investigations [4].

The survey also identified that smartphones contain valuable data that is not analyzed automatically because forensic tools do not support less popular apps. Some crimes are committed using specific types of apps. For example, stolen credit card information is used to purchase goods from online shops that are sent to dummy addresses. Online shopping and package tracking apps would contain valuable information in such investigations, but forensic professionals have to search such apps manually because they are not supported by forensic tools.

Another gap in contemporary forensic tools involves their handling of foreign apps. The survey indicated that the absence of documentation about foreign app functionality is problematic. Additionally, manual translation of the extracted app information from foreign languages can be expensive and leads to delays because information can be labeled only after it is translated.

From a technical point of view, the survey indicated that the extraction of data from smartphones continues to be problematic. The retrieval of user credentials from internal smartphone databases would be useful in investigations. It would also be useful to obtain deleted SQL entries and encrypted files. Recent research has demonstrated advances in restoring deleted SQLite entries [14].

5.2 Data Content Comparison

This section compares the representative smartphone content created as a result of the survey against the content of a published smartphone dataset analyzed by Goncalves et al. [7]. The published smartphone dataset content covers the same 11 data categories as the representative smartphone content.

The log-scale bar chart in Figure 1 shows the means of the data categories in the representative smartphone and in the published smartphone dataset. The lighter bars correspond to the representative smartphone whereas the darker bars correspond to the published dataset. The number above each data category is the variance corresponding to the representative smartphone relative to the mean distribution in the published dataset. Specifically, a value lower than one indicates that the published

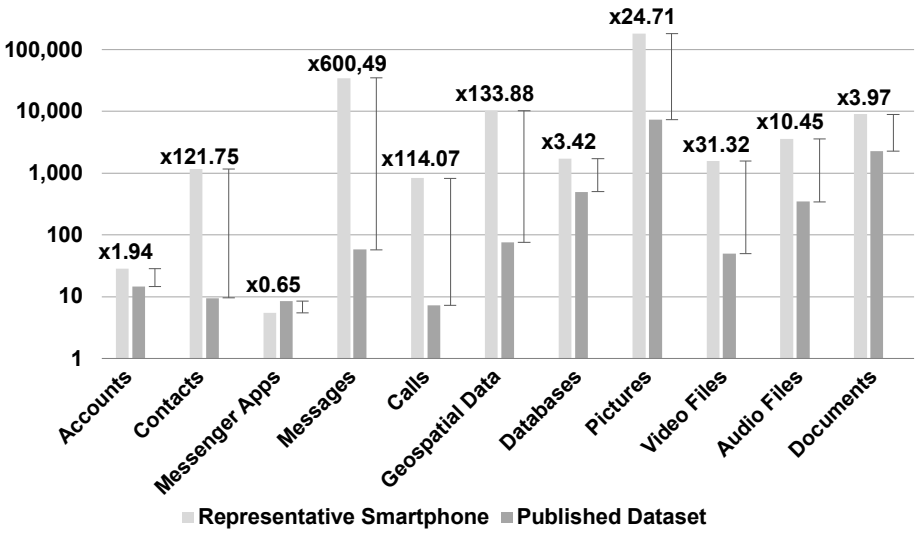

Figure 1. Representative smartphone content versus published dataset content.

dataset contains more files or entries in the particular data category compared with the representative smartphone whereas a value greater than one indicates that the published dataset contains less files or entries compared with the representative smartphone.

The number of messenger apps is lower by about one-third for the representative smartphone compared with the published dataset. For the other data categories, the representative smartphone has more content than the published dataset by factors ranging from two to about 600. Specifically, the representative smartphone contains about twice the number of accounts, 122 times more stored contacts, 600 times more text/email messages, 114 times more calls, 134 times more geospatial data, three times more database files, 25 times more picture files, 31 times more video files, ten times more audio files and four times more text documents than in the published dataset.

The amount of content corresponding to each data category in the published dataset was rated according to the minimum and maximum values of the representative smartphone shown in Table 1. Table 3 shows the published dataset content ratings. A data category in the published dataset is rated low (respectively, high) if its content is less than the minimum value (respectively, higher than the maximum value) of the representative smartphone. The data category in the published dataset is rated good if its content is within the minimum and maximum values of the representative smartphone.

Table 3. Published dataset content ratings based on a representative smartphone.

Low Rating	**Good Rating**	**High Rating**
Contacts	Accounts	Messenger apps
Calls	Text/email messages	
Geospatial data	Databases	
Pictures	Video files	
Documents	Audio files	

Only the number of messenger apps in the published dataset is greater than the number in the representative smartphone whose content corresponds to the contents of a real smartphone. Accounts, text/email messages, databases, video and audio files in the published dataset are rated good, although their numbers are very low compared with the representative smartphone. The remaining data categories have low ratings and, therefore, do not constitute realistic representations of real smartphones. The key finding is that the public dataset does not capture the complexity of real devices, which brings into question the realism of the published dataset and the quality of the forensic tools validated using the published dataset and other similar datasets.

The most notable difference between public and real datasets are the numbers of files and their contents. A possible explanation is that real devices are not only used to perform specific (criminal) acts, but are also constantly used in daily activities. This generates greater numbers of entries in the various smartphone databases which, in turn, results in forensic professionals encountering more files in smartphones. In contrast, public datasets are typically generated based on the specific scopes of the experiments instead of realistic user behavior.

6. Conclusions

The survey study of law enforcement professionals from Germany and Switzerland provides valuable information about the data in a representative smartphone encountered in a criminal investigation. Comparison of the distributions of data types in the representative smartphone against those in a published smartphone dataset revealed that real smartphones contain much more data than published datasets, which may be considered to be not very realistic; this calls into question validations of smartphone forensic techniques and tools based on the published datasets. The data distributions and the subsequent labeling of smartphone content as task-relevant and task-irrelevant assist researchers in creating more realistic datasets.

The survey reveals that the problems encountered by law enforcement professionals are similar to those identified in previous studies. Specifically, problems that persist include the lack of standardized forensic file formats and tool interoperability, bugs in forensic tools and long update cycles and absence of tool support for efficiently identifying relevant information. From the technical perspective, effective techniques and tools must be developed to access locked devices and encrypted content; an alternative solution is to encourage companies to install backdoor capabilities for law enforcement agencies. Forensic tool development efforts should focus on reducing bugs and vendors should provide tool support and release updates and patches on good schedules. Additionally, techniques should be developed to combat anti-forensic approaches that are increasingly being implemented in seized devices.

At this time, law enforcement professionals employ forensic techniques and tools that do not meet strict forensic examination requirements, and they often have to manually search for relevant information in seized devices. It is imperative that the digital forensics community institutes collaborative efforts to develop efficient techniques and cutting-edge tools as well as realistic forensic corpora that can help validate that the techniques and tools and the evidence proffered in court meet the highest standards.

References

[1] S. Abt and H. Baier, Are we missing labels? A study of the availability of ground truth in network security research, *Proceedings of the Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, pp. 40–55, 2014.

[2] K. Barmpatsalou, T. Cruz, E. Monteiro and P. Simoes, Current and future trends in mobile device forensics, *ACM Computing Surveys*, vol. 51(3), article no. 46, 2018.

[3] J. Camacho, K. Campos, P. Cedillo, B. Coronel and A. Bermeo, Forensic analysis of mobile devices: A systematic mapping study, in *Information and Communication Technologies of Ecuador*, M. Botto-Tobar, L. Barba-Maggi, J. Gonzalez-Huerta, P. Villacres-Cevallos, O. Gomez and M. Uvidia-Fassler (Eds.), Springer, Cham, Switzerland, pp. 57–72, 2018.

[4] E. Casey and C. Rose, Forensic analysis, in *Handbook of Digital Forensics and Investigation*, E. Casey (Ed.), Elsevier, Burlington, Massachusetts, pp. 21–62, 2010.

[5] European Parliament and Council, Regulation (EU) 2016/679, *Official Journal of the European Union*, vol. 59(L 119), pp. 1–88, 2016.

[6] S. Garfinkel, Digital forensics research: The next 10 years, *Digital Investigation*, vol. 7(S), pp. S64–S73, 2010.

[7] P. Goncalves, K. Dolovs, M. Stebner, A. Attenberger and H. Baier, Revisiting the Dataset Gap Problem – On Availability, Assessment and Perspectives of Mobile Forensic Corpora, Unpublished Manuscript, Cyber Defense Research Institute, Bundeswehr University, Munich, Germany, 2021.

[8] C. Grajeda, F. Breitinger and I. Baggili, Availability of datasets for digital forensics – And what is missing, *Digital Investigation*, vol. 22(S), pp. S94–S105, 2017.

[9] D. Lillis, B. Becker, T. O'Sullivan and M. Scanlon, Current challenges and future research areas for digital forensic investigations, *Proceedings of the Eleventh Annual Conference on Digital Forensics, Security and Law*, 2016.

[10] X. Lin, Chapter 9: File carving, in *Introductory Computer Forensics*, Springer, Cham, Switzerland, pp. 211–233, 2018.

[11] L. Luciano, I. Baggili, M. Topor, P. Casey and F. Breitinger, Digital forensics in the next five years, *Proceedings of the Thirteenth International Conference on Availability, Reliability and Security*, article no. 46, 2018.

[12] M. Meuser and U. Nagel, The expert interview and changes in knowledge production, in *Interviewing Experts*, A. Bogner, B. Littig and W. Menz (Eds.), Palgrave Macmillan, London, United Kingdom, pp. 17–42, 2009.

[13] A. Mylonas, V. Meletiadis, B. Tsoumas, L. Mitrou and D. Gritzalis, Smartphone forensics: A proactive investigation scheme for evidence acquisition, in *Information Privacy and Research*, D. Gritzalis, S. Furnell and M. Theoharidou (Eds.), Springer, Berlin Heidelberg, Germany, pp. 249–260, 2012.

[14] D. Pawlaszczyk and C. Hummert, Making the invisible visible – Techniques for recovering deleted SQLite data records, *International Journal of Cyber Forensics and Advanced Threat Investigations*, vol. 1(1-3), pp. 27–41, 2021.

[15] K. Woods, C. Lee, S. Garfinkel, D. Dittrich, A. Russell and K. Kearton, Creating realistic corpora for security and forensic education, *Proceedings of the Sixth Annual Conference on Digital Forensics, Security and Law*, 2011.