1 Literature review 2 - Digital forensic related datasets

1.1 Purpose of the literature review

Identify and summarize publicly available datasets that relates to digital forensic and consider their applicability for this thesis experiments. Table 1 shows examples of relevant datasets:

Catagory	Abbreviation	Example dataset
Forensics images	IMG	The Real Data Corpus (RDC)
Files	FILE	RAISE (RAw ImageS datasEt)
RAM dumps	RAM	
Network files	NET	
Malware	MAL	Kharon dataset
Email	EM	The Webb Spam Corpus 2011
SMS	SMS	
Password	PASS	Yahoo Password Frequency
Phishing	PHI	
Spam	SPAM	
Authorship	AUTH	Personae
Financial data/ fraud	FIN	
Forgery corpus	FORG	MICC-F2000

Table 1: Example of datasets

Decisions was made to limit the scope of the data collection, by excluding biometric datasets such as images of fingerprints, hand signature, gait, voice recognition and iris. But the review will include authorship attribution corpus.

1.2 Protocol/methodology

1. Search digital libraries and scan scientific articles for names, direct links or sources related to the datasets above and use this information on google search engine to identify individual datasets or repositories of datasets.
2. Document search phrases that resulted in identifying new datasets.
3. Repeat step 1 and 2 with other resources like github, keegle and figshare to locate more datsets.

1.3 Search phrases and justification

Documents was excluded from consideration if their title had little relation to information security, and if the document format was not easily searchable. An example of the latter case is pdf documents scanned by a scanner machine, where full text search of the text content is not applicable. Without the assistance of search, the process of finding the datasets would be too time consuming.

In table 1.3 is a summary of the collection phase of the literature review. Entries included in this table all lead to finding new datasets. An entry has an ID number, search phrase + search options, database name (search resource) and the number of hits for the search phrase. Entries with ID 1-5 is essentially full text search (matching based on meta data and text content). Fulltext search lead to more false positives, then only meta search. But was used in cases where the number of hits was manageable. An example for when fulltext was deemed unmanageable can be seen in entry 6, where meta search was used instead. The phrase 'forensic dataset' was used to find different types of relevant datasets, but this phrase alone is not good enough. This is because relevant papers may use publicly available datasets, but does not contain the term 'forensic'. Therefore more specific search terms from list in subsection 1.1 was also used. In entry 9 the NOT operator was used to discard biometric datasets. Entry 10 in table 1.3 returned hits that both included the phrase 'IDS dataset' and the term 'Network' in the meta data, and excluded hits that contained some already known network datasets. The term IDS was used to reduce the number of non-network related articles. This term may exclude some relevant hits, but its usage is justified as the other search phrases also covered some network related datasets. In entry 16 the first 10 results was used on Google to look find datasets on Github. This was done as it was tricky to identify relevant repositories using Githubs internal search. In entry 18 figshare did not provide the number of hits. Therefore Not Available (N/A) is in the #Hits column for this entry.

ID	Search phrase (comma (,) separates search options)	DB	#Hits
1	forensic corpora, exact phrase match	[a]	22
2	forensic corpus’, advanced search, both words must match (be present) in any field	[b]	9
3	forensic corpora’, advanced search, both words must match (be present) in any field	[b]	3
4	forensic dataset’, advanced search, both words must match (be present) in any field	[b]	61
5	forensic corpus, full text search	[c]	112
6	forensic dataset, in metadata only	[d]	94
7	malware dataset, in metadata only	[d]	174
8	((password dataset) NOT biometrics), in metadata only	[d]	19
9	Spam dataset, in metadata only	[d]	173
10	(((((((((IDS dataset) AND Network) NOT DARPA) NOT KDD) NOT KDD99cup) NOT DARPA98) NOT DARPA99) NOT DARPA-98) NOT DARPA-99) NOT NSL-KDD), in metadata only	[d]	118
11	fraud dataset, in metadata only	[d]	104
12	Forensic dataset, in All Sources(Computer Science), no books	[e]	1100
13	fraud	[f]	10
14	spam	[f]	3
15	email	[f]	18
16	dataset github	[g]	576000
17	spam	[h]	107
18	network	[h]	N/A

---

[a] https://link.springer.com/
[b] http://dl.acm.org/
[c] http://search.arxiv.org
[d] http://ieeexplore.ieee.org/
[e] http://www.sciencedirect.com
[f] https://www.kaggle.com
[g] https://www.google.no/
[h] https://figshare.com/

Table 2: Search summary

1.4 Search summary - datasets:

During the collection phase of the literature review, two related reviews was identified. The first review was from 2014 and identified 7 datasets[1]. The second review is as recent as 2017 and compiled a list online of 79 digital forensic related datasets [2],[3]. This review expands on the two reviews and its findings where largely independent from the two previous works.

Table 3 is a summary of the identified datasets in this review. An entry in this table is explained in the list below:

- Column Item = Numbered Item.
- Column C = Contribution, where S=dataset was obtained by the aid of supervisors, I=Thesis author found the same dataset independently from

the two reviews [1, 2], R=The reviews[1, 2] identified datasets that was not obtained by this review, N=This review identified datasets not present in [1, 2].

- Column Acc = Access, where P=public and R=By request
- Column DT = Data type, where S = Synthetic, R=Real and H=Hybrid Column CAT= Catagory, where the catagories abbriviations is shown in table 1
- Column Size= Size is either given in S=samples, GigaBytes (comrpressed/uncompressed) or Not avaliable (N/A)
- Column Description= A description that will include the name of the dataset, where it can be downloaded from, include original paper if available and additional details about the dataset.

Table 3: Datasets

I	C	Acc	DT	Cat	Size	Description
1	I	P	R	IMG	14 S	A collection of forensic images made/hosted by Brian Carrier[4]. The 14 forensic images can be divided up into the following categories: NTFS file systems, FAT file system, ISO9660 file system and a memory image. Brian created scenarios to test string search, partitions with multiple file systems, file carving etc.
2	I	R	R	IMG	70TB Compressed	The Real Data Corpus (RDC) is data collected of digital devices from the secondary market[5]. The dataset contains hard drives images, flash memory images and CDROMS. According
3	I	P	S	IMG	16 S	Computer Forensic Reference Data Sets (CFReDS) can be used for forensic tool testing[6]. CFReDS includes forensic images and simulated data for memory forensics, file carving, string search and file recovery.
4	I	R	R	AUTH	609 S	Polish Corpus of Suicide Notes (PCSN) are real suicide letter written by both young and old polish men and women from the period of 1999-2009[7],[8].
5	I	P	R	AUTH	12 S	The Brennan-Greenstadt corpus contains two documents from each of the 12 participating authors[9], [10]. In the first text the authors attempted to obfuscate the characteristics of their writing. And in the second text the authors tried to imitate the writing style of a different writer.

. . . continued

I	C	Acc	DT	Cat	Size	Description
6	I	R	R	AUTH	145 S	The paper claims that the size of the German corpus Personae makes it possible to classify the author of the text as well as the author personality[11],[12]. Personae consist of 145 bachelor student essays with lengths around 1400 words. The students, took a personality test. This test made classification of their personality possible. But it is difficult to infer from the sources [11],[12] whether the personality test is part of the dataset or not.
7	I	P	R	AUTH	12338 S	This corpus is a subset of the Enron dataset and can be used for authorship attribution and verification. 24% of the samples is from non-Enron authors while the rest is from the Enron set[13],[14]. Names and email addresses was omitted from the dataset.
8	I	P	R	AUTH	N/A	The dataset contains training and test data for several authorship attribution scenarios based on works of fiction. Each scenario has a different amount of authors, number of documents, and minimum word length[15],[14].
9	I	P	R	AUTH	110 S	Authorship classification on English, Spanish and Greek texts. Most of the documents are in the word length range 1001-1500 words[16],[14] .
10	I	P	R	AUTH	4959 S	Authorship attribution corpus with documents written in English, Dutch, Spanish, and Greek[17], [14]. University students created the Dutch and English documents. And the Spanish and Greek documents was obtained from newspapers.
11	I	P	R	AUTH	3701 S	Authorship attribution corpus with documents written in English, Dutch, Spanish, and Greek. The authors of the Dutch documents was Students at a university in Belgium[18],[14]. English documents was taken from theatre plays. Spanish and Greek documents was obtained from opinion articles.
12	I	P	R	AUTH	1000 S	Reddit Cross-Topic AV Corpus consist of 1000 reddit users and their comments from 2010-2016 on 1388 different subjects [19], [20].

. . . continued

I	C	Acc	DT	Cat	Size	Description
13	I	P	R	FILE	350GB N/A	RAISE (RAw ImageS datasEt): 8156 Unprocessed and high resolution images. The images are taken by the following cameras: Nikon D40, Nikon D90 and Nikon D7000[21], [22]. The original paper states that this dataset can be useful to test image forgery algorithms[21].
14	I	P	S	NET	38/50 S	DARPA 1998 and 1999 is datasets of simulated network traffic used to assess the detection capabilities of intrusion detection systems[23],[24]. DARPA 1998 contains 38 categories of UNIX based attacks. DARPA 1999 increases the number of categories to 50 and added Windows NT based exploits as well.
15	I	P	S	NET	2 S	DARPA 2000 has simulated data from two distributed denial of service attacks[25].
16	I	P	R	EM	5000000 S	The Global Intelligence files (GIfiles) are a collection of 5 million leaked emails from Stratfor, that gives insight into how the intelligence community operates[26], [27].
17	I	R	R	FILE	10 billion S	SherLock is a Android Smartphone dataset that contains running application/process information, sensory data and OS data captured with normal user privileges[28], [29]. The dataset also have labels that can be assign to describe ongoing malicious activity on the phone.
18	I	R	R	MAL	29385674 S	VirusShare.com is a virus sharing website with currently 29385674 malware samples [30].
19	I	P	R	FORG	220/2000 S	MICC-F220 and MICC-F2000 are datasets that contains untouched images and images where parts of the image is modified by scaling, rotating and scaling[31], [32]. The datasets have been used to benchmark a copy-move forgery algorithm.
20	I	P	R	MAL	$\approx 500GB$	A dataset for classifying known malware and their associated malware family[33]. There are in total 500GB worth of malware samples, that belongs into one of 9 families of malware.

. . . continued

I	C	Acc	DT	Cat	Size	Description
21	I	R	R	MAL	5560 S	The Drebin Dataset have 5560 malicious android applications that can be categorized into one of 179 malware families[34], [35].
22	I	P	R	SMS	87300 S	The NUS SMS Corpus includes 55835 English and 31465 Chinese SMS messages[36], [37]. To avoid bias or promote message diversity in the sampling process, the individual SMS messages was captured without considering any particular topic. The SMS messages can be download in JSON, XML and SQL format.
23	I	P	R	AUTH	19320 S	A authorship corpus of 681288 Blog entries and 19320 problems[38],[39]
24	I	P	S	AUTH	20 S	A capture the flag (CTF) authorship corpus[40],[41]. The corpus have been used in the multi-classification problem of classifying the origin of the exploit attempts to one of 20 CTF teams. The data is available in JSON format and includes source and destination of attack, timing information and histogram of payload.
25	I	P	R	SPAM	≈ 350000 S/ 1GB compressed	The Webb Spam Corpus 2011: A custom crawler was built to collect spam web pages[42], [43]. The resulting collection was preprocessed to remove instances of legitimate websites and websites that could not get resolved. The dataset contains both the spam and the HTTP sessions for the spam servers.
26	I	P	R	PASS	N/A	Yahoo Password Frequency Corpus: A sanitized password frequency corpus that protect the privacy of the user accounts[44], [45] . The scheme also protects up to two duplicate accounts, that has similar passwords. The sanitization is performed to prevent adversaries to gain knowledge of individual users.
27	I	P	S	MAL	399 S	DroidWare is a malware dataset for the android platform. The dataset is made up of 278 benign and 121 malicious samples[46],[47]. Each sample has a 152 feature vector of Android application permissions.
28	I	P	S	MAL	4S	Synthetic dataset with 4 botnet samples. The botnet actions in each sample differs from injection, reconnaissance, command and control (C&C) communication channels and botnet prorogation[48],[49].

. . . continued

I	C	Acc	DT	Cat	Size	Description
28	I	P	R	MAL	7 S	Kharon dataset contains malware documentation, that has been used to benchmark GroddDroid capability to trigger malicious code[50], [51]. The documentation was obtained though Static and dynamic analysis on a set of malware samples. The documentation includes the location of the malicious code blocks, the trigger conditions, and how the malware acts when triggered.
29	I	P	R	NET	N/A	The MAWILab database contains labels, that categorize network anomalies. It can be used to assist in evaluating the performance of intrusion detection systems (IDS)[52],[53].
30	I	P	S	NET	743M	KDD Cup 1999 Data: A synthetic dataset that is made up off network traffic samples[54]. These samples is labelled benign or malign[55]. Malign samples are attempting to attack availability, to perform privilege escalation, to imitating a local user and to perform reconnaissance. The dataset is produced based on the DARPA98 dataset.
31	I	P	H	NET	2540044 S	UNSW-NB15: The samples are labelled malign or benign. Each sample has 49 features that includes variables such as time to live (TTL), IP information, sequence number, time between TCP SYN and TCP ACK etc[56], [57]. The malicious samples aims to identify vulnerabilities by perform active reconnaissance and by using fuzzed inputs. The malware samples also attempts to install backdoors, target the availability of services, opening a shell to run arbitrary code and to compromise new hosts. There are in total 2 540 044 samples spread across 4 .csv files, a smaller subset of this dataset is used to create a training and a test set.

. . . continued

I	C	Acc	DT	Cat	Size	Description
32	I	R	R	FILE	5546565 S	AndroZoo dataset includes over 5 million android applications (APKs)[58],[59]. The APKs was obtained by crawling multiple APKs distributors such as google play, AppChina, torrents etc. Efforts was made to avoid downloading duplicate files from the same vendor. But creators of the dataset gives no guarantees that the same file was not downloaded from multiple vendors. Each sample contains a zipped apk file, with its byte code, meta data, signed certificate and miscellaneous files.
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						
28						

. . . continued

I	C	Acc	DT	Cat	Size	Description
28						
28						
28						
28						
28						
28						
28						
28						

Table 3: Datasets

References

[1] Yannikos Y, Graner L, Steinebach M, Winter C. In: Peterson G, Shenoi S, editors. Data Corpora for Digital Forensics Education and Research. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014. p. 309–325. Available from: `https://doi.org/10.1007/978-3-662-44952-3_21`.

[2] Grajeda C, Breitinger F, Baggili I. Availability of datasets for digital forensics. And what is missing. Digital Investigation. 2017;22(Supplement):S94 – S105. Available from: `http://www.sciencedirect.com/science/article/pii/S1742287617301913`.

[3] Grajeda C, Breitinger F, Baggili I. DATASETS FOR CYBER FORENSICS; 2017. Last accessed (DD/MM/YYYY) 19/09/2017. Available from: `http://datasets.fbreitinger.de/datasets/`.

[4] Carrier B. Digital Forensics Tool Testing Images; 2010. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `http://dftt.sourceforge.net/`.

[5] Corpora D. Real Data Corpus; 2017. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `https://digitalcorpora.org/corpora/disk-images/real-data-corpus`.

[6] NIST. The CFReDS Project; 2016. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `https://www.cfreds.nist.gov/`.

[7] Marcinczuk M, Zasko-Zielinska M, Piasecki M. Structure Annotation in the Polish Corpus of Suicide Notes. In: Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings; 2011. p. 419–426. Available from: `https://doi.org/10.1007/978-3-642-23538-2_53`.

[8] Zaśko-Zielińska M, Piasecki M, Marcińczuk M. Polski korpus listów pożegnalnych samobójców; 2015. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `http://www.pcsn.uni.wroc.pl/`.

[9] Brennan M, Greenstadt R. Practical Attacks Against Authorship Recognition Techniques. In: Proceedings of the 21st Innovative Applica-

tions of Artificial Intelligence Conference, IAAI-09; 2009. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `https://www.cs.drexel.edu/~greenie/brennan_paper.pdf`.

[10] psal cs drexel edu. JStylo-Anonymouth; 2013. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `https://psal.cs.drexel.edu/index.php/JStylo-Anonymouth`.

[11] Luyckx K, Daelemans W. Personae: a Corpus for Author and Personality Prediction from Text. In: LREC. European Language Resources Association; 2008. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `http://www.lrec-conf.org/proceedings/lrec2008/pdf/759_paper.pdf`.

[12] Luyckx K, Daelemans W. Personae Corpus; 2017. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `https://www.clips.uantwerpen.be/datasets/personae-corpus`.

[13] Argamon S, Juola P. Overview of the International Authorship Identification Competition at PAN-2011. In: Petras V, Forner P, Clough P, editors. Notebook Papers of CLEF 2011 Labs and Workshops, 19-22 September, Amsterdam, Netherlands; 2011. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `http://www.uni-weimar.de/medien/webis/events/pan-11/pan11-papers-final/pan11-author-identification/argamon11-overview.pdf`.

[14] pan webis de. Evaluation Data; 2016. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `http://pan.webis.de/data.html`.

[15] Juola P. An Overview of the Traditional Authorship Attribution Subtask. In: Forner P, Karlgren J, Womser-Hacker C, editors. CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20 September, Rome, Italy; 2012. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `http://www.uni-weimar.de/medien/webis/events/pan-12/pan12-papers-final/pan12-author-identification/juola12-overview.pdf`.

[16] Juola P, Stamatatos E. Overview of the Author Identification Task at PAN 2013. In: Forner P, Navigli R, Tufis D, editors. CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain; 2013. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `http://www.uni-weimar.de/medien/webis/events/pan-13/pan13-papers-final/pan13-authorship-verification/juola13-overview.pdf`.

[17] Stamatatos E, Daelemans W, Verhoeven B, Potthast M, Stein B, Juola P, et al. Overview of the Author Identification Task at PAN 2014. In: Cappellato L, Ferro N, Halvey M, Kraaij W, editors. CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK. CEUR-WS.org; 2014. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `http://www.uni-weimar.de/medien/webis/events/pan-14/pan14-papers-final/pan14-authorship-verification/stamatatos14-overview.pdf`.

[18] Stamatatos E, amd Ben Verhoeven WD, Juola P, López-López A, Potthast M, Stein B. Overview of the Author Identification Task at

PAN 2015. In: Cappellato L, Ferro N, Jones G, San Juan E, editors. CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France. CEUR-WS.org; 2015. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-papers-final/pan15-authorship-verification/stamatatos15-overview.pdf`.

[19] Halvani O, Winter C, Graner L. On the Usefulness of Compression Models for Authorship Verification. In: Proceedings of the 12th International Conference on Availability, Reliability and Security. ARES '17. New York, NY, USA: ACM; 2017. p. 54:1–54:10. Available from: `http://doi.acm.org/10.1145/3098954.3104050`.

[20] Halvani O, Winter C, Graner L. ARES_WSDF2017; 2017. Last accessed (DD/MM/YYYY) 20/09/2017. Available from: `https://www.dropbox.com/sh/f2mlp6u5vervx9b/AABr_c7qrmahCqUviIu3ORz6a?dl=0`.

[21] Dang-Nguyen DT, Pasquini C, Conotter V, Boato G. RAISE: A Raw Images Dataset for Digital Image Forensics. In: Proceedings of the 6th ACM Multimedia Systems Conference. MMSys '15. New York, NY, USA: ACM; 2015. p. 219–224. Available from: `http://doi.acm.org/10.1145/2713168.2713194`.

[22] Dang-Nguyen DT, Pasquini C, Conotter V, Boato G. Introducing RAISE dataset;. Last accessed (DD/MM/YYYY) 21/09/2017. Available from: `http://mmlab.science.unitn.it/RAISE/`.

[23] ll mit edu. DARPA Intrusion Detection Data Sets;. Last accessed (DD/MM/YYYY) 21/09/2017. Available from: `https://ll.mit.edu/ideval/data/index.html`.

[24] Haines JW, Lippman RP, Fried DJ, Zissman MA, Tran E, Boswell SB. 1999 DARPA INTRUSION DETECTION EVALUATION DESIGN AND PROCEDURES; 2001. Last accessed (DD/MM/YYYY) 21/09/2017. Available from: `https://ll.mit.edu/ideval/files/TR-1062.pdf`.

[25] ll mit edu. 2000 DARPA Intrusion Detection Scenario Specific Data Sets;. Last accessed (DD/MM/YYYY) 21/09/2017. Available from: `https://ll.mit.edu/ideval/data/2000data.html`.

[26] Harrison S. The Global Inteligence Files;. Last accessed (DD/MM/YYYY) 21/09/2017. Available from: `https://wikileaks.org/the-gifiles.html`.

[27] wlstorage net. Index of /torrent/gifiles/;. Last accessed (DD/MM/YYYY) 21/09/2017. Available from: `https://wlstorage.net/torrent/gifiles/`.

[28] Mirsky Y, Shabtai A, Rokach L, Shapira B, Elovici Y. SherLock vs Moriarty: A Smartphone Dataset for Cybersecurity Research. In: Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security. AISec '16. New York, NY, USA: ACM; 2016. p. 1–12. Available from: `http://doi.acm.org/10.1145/2996758.2996764`.

[29] Mirsky Y, Shabtai A, Rokach L, Shapira B, Elovici Y. DOWNLOADS; 2016. Last accessed (DD/MM/YYYY) 21/09/2017. Available from: `http://bigdata.ise.bgu.ac.il/sherlock/#/download`.

[30] VirusShare. VirusShare.com - Because Sharing is Caring; 2017. Last accessed (DD/MM/YYYY) 21/09/2017. Available from: `https://virusshare.com/`.

[31] Amerini I, Ballan L, Caldelli R, Del Bimbo A, Serra G. A SIFT-Based Forensic Method for Copy&#x2013;Move Attack Detection and Transformation Recovery. Trans Info For Sec. 2011 Sep;6(3):1099–1110. Available from: `http://dx.doi.org/10.1109/TIFS.2011.2129512`.

[32] lambertoballan. sift-forensic; 2015. Last accessed (DD/MM/YYYY) 21/09/2017. Available from: `https://github.com/lambertoballan/sift-forensic/blob/master/README.md`.

[33] Kaggle. Microsoft Malware Classification Challenge (BIG 2015);. Last accessed (DD/MM/YYYY) 21/09/2017. Available from: `https://www.kaggle.com/c/malware-classification/data`.

[34] Arp D, Spreitzenbarth M, Gascon H, Rieck K. Drebin: Effective and explainable detection of android malware in your pocket; 2014. Last accessed (DD/MM/YYYY) 21/09/2017. Available from: `https://www.tu-braunschweig.de/Medien-DB/sec/pubs/2014-ndss.pdf`.

[35] Arp D, Spreitzenbarth M, Gascon H, Rieck K. The Drebin Dataset; 2016. Last accessed (DD/MM/YYYY) 21/09/2017. Available from: `https://www.sec.cs.tu-bs.de/~danarp/drebin/`.

[36] Chen T, Kan MY. Creating a live, public short message service corpus: the NUS SMS corpus. Language Resources and Evaluation. 2013 Jun;47(2):299–335. Available from: `https://doi.org/10.1007/s10579-012-9197-9`.

[37] kite1988. nus-sms-corpus; 2016. Last accessed (DD/MM/YYYY) 22/09/2017. Available from: `https://github.com/kite1988/nus-sms-corpus`.

[38] Schler J, Koppel M, Argamon S, Pennebaker J. In: Effects of age and gender on blogging. vol. SS-06-03; 2006. p. 191–197. Last accessed (DD/MM/YYYY) 22/09/2017. Available from: `http://u.cs.biu.ac.il/~schlerj/schler_springsymp06.pdf`.

[39] u cs biu ac il. The Blog Authorship Corpus;. Last accessed (DD/MM/YYYY) 22/09/2017. Available from: `http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm`.

[40] Nunes E, Shakarian P, Simari GI, Ruef A. Argumentation Models for Cyber Attribution. CoRR. 2016;abs/1607.02171. Last accessed (DD/MM/YYYY) 22/09/2017. Available from: `https://arxiv.org/pdf/1607.02171v1.pdf`.

[41] Nunes E. CTF data;. Last accessed (DD/MM/YYYY) 22/09/2017. Available from: `https://www.dropbox.com/sh/17d4eyg0cwoxg8s/AAA5g1NvQw-tUoZPvldloddRa?dl=0`.

[42] Wang D, Irani D, Pu C. Evolutionary Study of Web Spam: Webb Spam Corpus 2011 Versus Webb Spam Corpus 2006. In: Proceedings of the 2012 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2012). COLLABORATECOM '12. Washington, DC, USA: IEEE Computer Society; 2012. p. 40–49. Last accessed (DD/MM/YYYY) 22/09/2017. Available from: `http://de-wang.org/download/webbspamcorpus2011.pdf`.

[43] Wang D, Irani D, Pu C. Webb Spam Corpus 2011;. Last accessed (DD/MM/YYYY) 22/09/2017. Available from: `https://www.cc.gatech.edu/projects/doi/WebbSpamCorpus.html`.

[44] Bonneau J. Yahoo Password Frequency Corpus. 2015 12;Available from: `https://figshare.com/articles/Yahoo_Password_Frequency_Corpus/2057937`.

[45] Blocki J, Datta A, Bonneau J. Differentially Private Password Frequency Lists. IACR Cryptology ePrint Archive. 2016;2016:153. Available from: `http://eprint.iacr.org/2016/153`.

[46] d Costa KAP, d Silva LA, Martins GB, Rosa GH, Pereira CR, Papa JP. Malware Detection in Android-Based Mobile Environments Using Optimum-Path Forest. In: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA); 2015. p. 754–759.

[47] RECOVI. DroidWare; 2015. Last accessed (DD/MM/YYYY) 23/09/2017. Available from: `https://github.com/RECOVI/DroidWare`.

[48] Bowen T, Poylisher A, Serban C, Chadha R, Chiang CYJ, Marvel LM. Enabling reproducible cyber research - four labeled datasets. In: MILCOM 2016 - 2016 IEEE Military Communications Conference; 2016. p. 539–544.

[49] McDaniel P. Data Sets;. Last accessed (DD/MM/YYYY) 23/09/2017. Available from: `https://cybervan.appcomsci.com:9000/datasets`.

[50] Kiss N, Lalande JF, Leslous M, Tong VVT. Kharon Dataset: Android Malware under a Microscope. In: The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2016). San Jose, CA: USENIX Association; 2016. p. 1–12. Available from: `https://www.usenix.org/conference/laser2016/program/presentation/kiss`.

[51] Kharon-project. Kharon Malware Dataset; 2016. Last accessed (DD/MM/YYYY) 23/09/2017. Available from: `http://kharon.gforge.inria.fr/dataset/`.

[52] ;. .

[53] fukuda lab. MAWILab v1.1; 2017. Last accessed (DD/MM/YYYY) 23/09/2017. Available from: `http://www.fukuda-lab.org/mawilab/data.html`.

[54] "kdd ics uci edu". KDD Cup 1999 Data; 1999. Last accessed (DD/MM/YYYY) 23/09/2017. Available from: `http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html`.

[55] Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A Detailed Analysis of the KDD CUP 99 Data Set. In: Proceedings of the Second IEEE International Conference on Computational Intelligence for Security and Defense Applications. CISDA'09. Piscataway, NJ, USA: IEEE Press; 2009. p. 53–58. Available from: `http://dl.acm.org/citation.cfm?id=1736481.1736489`.

[56] Moustafa N, Slay J. The Significant Features of the UNSW-NB15 and the KDD99 Data Sets for Network Intrusion Detection Systems. In: 2015 4th International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS); 2015. p. 25–31.

[57] unsw-adfa-edu au". The UNSW-NB15 data set description; 2016. Last accessed (DD/MM/YYYY) 24/09/2017. Available from: `https://www.unsw.adfa.edu.au/australian-centre-for-cyber-security/cybersecurity/ADFA-NB15-Datasets/`.

[58] Allix K, Bissyandé TF, Klein J, Traon YL. AndroZoo: Collecting Millions of Android Apps for the Research Community. In: 2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR); 2016. p. 468–471.

[59] androzoo. androzoo; 2016. Last accessed (DD/MM/YYYY) 24/09/2017. Available from: `https://androzoo.uni.lu/`.