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 contents	RAM	Memory Buddies Traces(MBT)
Network	NET	Common crawl
Malware	MAL	Kharon dataset
Email	EM	The Webb Spam Corpus 2011
SMS	SMS	The NUS SMS Corpus
Password	PASS	Yahoo Password Frequency
Phishing	PHI	Phishing Websites Data
Spam	SPAM	TREC 2011
Authorship	AUTH	Personae
Financial data/ fraud	FIN	CMS dataset
Forgery corpus	FORG	MICC-F2000
Collection of different datasets	COLL	CAIDA data

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.

I	Name	Cat	This paper	in review[?]	in review[?]
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					
31					
32					
33					
34					
35					
36					
37					
38					
39					
40					
41					
42					
43					
44					
45					
46					
47					
48					
49					
50					
51					
52					
53					
54					
55					
56					
57					
58					
59					
60					
61					
62					
63					
64					
65					
66					
67					
68					
69					
70					
71					
72					
73					
74					
75					
76					
77					
78					
79					
80					
81					
82					
83					

Table 2: Where the datasets/collections can be found

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 3: 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 4 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 4: Datasets

I	Name	Acc	DT	Cat	Size	Description
1	BAC	P	R	AUTH	19320 S	The Blog Authorship Corpus (BAC): A authorship corpus of 681288 Blog entries and 19320 problems[4],[5] Obtain dataset here
2	CTFAC	P	S	AUTH	20 S	A capture the flag (CTF) authorship corpus (CTFAC)[6],[7]. 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. Obtain dataset here
3	PCSN	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[8],[9]. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
4	TBGC	P	R	AUTH	12 S	The Brennan-Greenstadt corpus (TBGC) contains two documents from each of the 12 participating authors[10], [11]. 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. Obtain dataset here
5	Personae	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[12],[13]. 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 [12],[13] whether the personality test is part of the dataset or not. Obtain dataset here
6	PAN-Enron	P	R	AUTH	12338 S	PAN-Enron 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[14],[15]. Names and email addresses was omitted from the dataset. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
7	PAN/CLEF12	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[16],[15]. Obtain dataset here
8	PAN13	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[17],[15]. Obtain dataset here
9	PAN14	P	R	AUTH	4959 S	Authorship attribution corpus with documents written in English, Dutch, Spanish, and Greek[18], [15]. University students created the Dutch and English documents. And the Spanish and Greek documents was obtained from newspapers. Obtain dataset here
10	PAN15	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[19],[15]. English documents was taken from theatre plays. Spanish and Greek documents was obtained from opinion articles. Obtain dataset here
11	RCTAC	P	R	AUTH	1000 S	Reddit Cross-Topic AV Corpus (RCTAC) consist of 1000 reddit users and their comments from 2010-2016 on 1388 different subjects [20], [21]. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
12	MUD03	P	N/A	AUTH	750000 S	Masquerading User Data 2003 (MUD03). A dataset of 750000 UNIX commands[22], [23]. The commands are either from one of 50 legitimate users or a user imitating one of the 50. Each legitimate user has 5000 commands that is theirs and a random proportion of 10000 commands that is attribute to them or a masquerading person. Obtain dataset here
13	netresec	P	P,R	COLL	N/A	This website have compiled a list of available PCAP files online[24]. The list contain PCAP files that have been used in competitions, conferences, and PCAP files that contain malware or exploits. Obtain dataset here
14	MTA13-17	P	N/A	COLL	≈ 1100 S	malware-traffic-analysis (MTA) website host a collection of PCAP files and malware samples[25]. Obtain dataset here
15	pcapr	P	S,R	COLL	60507109 S_1 / 3465 S_2	A large publicly available searchable database that contains 60507109 packets and 3465 pcaps[26]. Obtain datasets here
16	PCAPsDB	P	S,R	COLL	N/A	PCAPsDB is a repository of different PCAP files, some of which are Malware[27]. Obtain dataset here
17	CAIDA	P	R	COLL	71 S	CAIDA Data: A collection that contains a mixture of publicly availably and by request network datasets[28]. These network dataset focuses on Autonomous System (AS) topology, denial of service attacks, worms, anonymized passive network traffic monitoring and darknet traffic. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
18	csmining	P	R	COLL	10 S	A collection of spam datasets, news articles, system calls from malware and Reuters-21578 dataset used for text categorization[29]. Obtain datasets here
19	AZSecure-data	P	R	COLL	111 S	Is a resource that contains list of geopolitical discussions, dark web forum threads, phising and legitimate sites, malware, network traffic and data from social media communication[30]. Obtain datasets here
20	Digital Corpora	P,R	R	COLL	N/A	Is a collection that contain: • Cell phone images • Disk images and the already mentioned RDC dataset. • Network traffic including the already mentioned DARPA 1998,1999 and 2000 datasets. • Govdocs1 a dataset with ≈ 1000000 files. [31] Obtain datasets here
21	DFCF review			COLL		DATASETS FOR CYBER FORENSICS (DFCF) - summerize short the collection of the datasets I do not have .... TODO.
22	GIfiles	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[32], [33]. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
23	USHCE	P	R	EM	60 MB S	≈ 7000 PDF pages of [US president candidate] Hillary Clinton emails (USHCE) was released as a result of a freedom of information claim[34]. The dataset contains both email content and metadata. Obtain dataset here
24	419 dataset	P	R	EM	2500 S	419 fraud emails: Content and metadata from 2500 fraud emails[35]. Obtain dataset here
25	MLE200	P	R	EM	200 S / ≈ 2MB	Multilanguage emails(MLE): Spanish, English Portuguese Emails[36]. Obtain dataset here
26	Enrondata	P	R	EM	N/A	Enrondata is a website that has a compiled list that links to differernt versions of the ENRON dataset that is in PST or MIME format and with different record sizes[37]. Obtain dataset here
27	RAISE	P	R	FILE	350GB N/A	RAw ImageS datasEt (RAISE): 8156 Unprocessed and high resolution images. The images are taken by the following cameras: Nikon D40, Nikon D90 and Nikon D7000[38], [39]. The original paper states that this dataset can be useful to test image forgery algorithms[38]. Obtain dataset here
28	SherLock	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[40], [41]. The dataset also have labels that can be assign to describe ongoing malicious activity on the phone. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
29	AndroZoo	R	R	FILE	5546565 S	AndroZoo dataset includes over 5 million android applications (APKs)[42],[43]. 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. Obtain dataset here
30	CTD15	P	R	FIN	68MB / 284807 S	Credit transaction dataset (CTD) A numerical dataset of 284807 bank transactions[44], [45]. Of all the transactions only 492 of them are fraudulent. In the feature vector there are 28 principal components, the elapsed time between transactions, the transactions amount, and the fraud/not fraud class label. Obtain dataset here
31	UCSD-FICO-09	P	R	FIN	$>$ 100000 S	UCSD-FICO-09 is a electronic commerce fraud dataset with anonymized features [46], [47]. The numerical dataset has both labelled (training) and unlabelled(testing) samples. Obtain dataset here
32	CMS	P	R	FIN	$\approx$ 6 GB	Real financial statements in .csv format from Centers for Medicare & Medicaid Services [US], in the year 2013-2016[48], [49]. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
33	PaySim	P	S	FIN	182MB	A smaller subset of the Synthetic PaySim dataset is available on Kaggle[50], [51]. The synthetic data was generated based on real world examples. The feature vector contain time information, transaction type, identifier for the user(s) involved in a transaction, old and current state of the balance sheet and class labels regarding if the transactions is considered fraudulent. Obtain dataset here
34	BankSim	P	S	FIN	13MB	BankSim is a synthetic fraud dataset with 587443 legitimate and 7200 illegitimate transactions/samples[52], [53]. The simulating agent generating the dataset is based on real bank transactions. The feature vector contains age range, gender catagory, what the transaction was spent on, zip code etc. Obtain dataset here
35	MICC	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[54], [55]. The datasets have been used to benchmark a copy-move forgery algorithm. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
36	Brian Carrier	P	R	IMG	14 S	A collection of forensic images made/hosted by Brian Carrier[56]. 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. Obtain dataset here
37	RDC	R	R	IMG	70TB Compressed	The Real Data Corpus (RDC) is data collected of digital devices from the secondary market[57]. The dataset contains hard drives images, flash memory images and CDROMS. According Obtain dataset here
38	CFReDS	P	S	IMG	16 S	Computer Forensic Reference Data Sets (CFReDS) can be used for forensic tool testing[58]. CFReDS includes forensic images and simulated data for memory forensics, file carving, string search and file recovery. Obtain dataset here
39	VirusShare	R	R	MAL	29385674 S	VirusShare.com is a virus sharing website with currently 29385674 malware samples [59]. Obtain dataset here
40	BIG15	P	R	MAL	$\approx 500GB$	A dataset for classifying known malware and their associated malware family[60]. There are in total 500GB worth of malware samples, that belongs into one of 9 families of malware. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
41	Drebin	R	R	MAL	5560 S	The Drebin Dataset have 5560 malicious android applications that can be categorized into one of 179 malware families[61], [62]. Obtain dataset here
42	DroidWare	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[63],[64]. Each sample has a 152 feature vector of Android application permissions. Obtain dataset here
43	MILCOM16	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[65],[66]. Obtain dataset here
44	Kharon	P	R	MAL	7 S	Kharon dataset contains malware documentation, that has been used to benchmark GroddDroid capability to trigger malicious code[67], [68]. 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. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
45	Mudflow	P	R	MAL	18204 S	The Mudflow dataset was used to train a ”benign or malign” binary classifier, on the information flow from benign Android applications obtained from the Google store[69], [70]. Instead of focusing on what resources the applications request access to, the classifier determines if the usage of these resources are to be deemed normal. The dataset contains 2866 benign and 15338 malicious apps. Obtain dataset and scripts here
46	ISOT2010	P	R	MAL	N/A	ISOT is a dataset that is built up of benign and malicious network traffic, from multiple sources[71], [72], [73]. The malicious samples is off the Storm and Waledac botnets. Obtain dataset here
47	ECML/PKDD07	R	H	MAL	50000 S	ECML/PKDD-2007: A dataset made up of 40 000 normal queries, 9000 exploits with descriptions on the target environment and 1000 exploits without target information [74]. The dataset is in XML format and contain attacks from 7 different categories. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
48	CSIC10	P	R	MAL	610000 S	The motivation behind the creation of the HTTP-CSIC-2010 dataset, was a shortage of malware datasets that exploits real web applications[75]. HTTP-CSIC-2010 contains malign and benign labelled request against a Spanish electronic commerce web application. The samples with malicious consequence target the confidentiality and integrity of the web application resources. Obtain dataset here
49	BlogPcap	P	N/A	MAL	1000 S	A list of 1000 Malware PCAP files. The blog also contain other malware samples as well[76], [77]. Obtain dataset here
50	Malrec	P	R	MAL	24389 S	The Malrec Dataset: A dataset of the system calls and arguments made by malicious software[78]. Obtain dataset her
51	CTU-13	P	R	MAL	334 S	Malware Capture Facility Project: A repository of Botnet and benign network traffic[79]. Obtain dataset here
52	ISCX Botnet	R	H	MAL	N/A	ISCX Botnet: Is a derivative dataset from multiple sources[80], [81]. The dataset contains normal traffic and malicious traffic from 16 different families of botnets. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
53	ISCXAB	R	R	MAL	1929 S	ISCX Android Botnet (ISCXAB): This dataset is built up of AnserverBot, Bmaster, DroidDream, Geinimi, MisoSMS, NickySpy, Not Compatible, PJapps, Pletor, RootSmart, Sandroid, TigerBot, Wroba and Zitmo botnets in the form of Android application package (APK) files[82], [83]. Obtain dataset here
54	DARPA98/99	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[84],[85]. 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. Obtain dataset here
55	DARPA2000	P	S	NET	2 S	DARPA 2000 has simulated data from two distributed denial of service attacks[86]. Obtain dataset here
56	MAWILab	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)[87],[88]. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
57	KDD Cup99	P	S	NET	743M	KDD Cup 1999 Data: A synthetic dataset that is made up off network traffic samples[89]. These samples is labelled benign or malign[90]. 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. Obtain dataset here
58	UNSW-NB15	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[91], [92]. 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. Obtain dataset here
59	NSA-CDX	P	R	NET	5 S	A collection of Cyber Defense Exercises (CDX) from the National Security Agency (NSA)[93]. In the CDX 2009 collection there are logs of DNS, web server, and IDS. Obtain datasets here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
60	ADFA	P	N/A	NET	N/A	The ADFA Intrusion Detection Datasets: According to the author in [94] the ADFA dataset has higher complexity, more attack options, frequency of attacks/normal traffic is more evenly distributed, and more extensive in scope then the KDD 98/99 evaluation datasets[95]. Given the reasons above the author concludes that the ADFA is the better dataset to assess the performance of Host based IDS (HIDS). Obtain datasets here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
61	Kyoto data	P	R	NET	19683 MB	A numerical dataset that tracks network activity from Honeypots and sensors that is in the management control of Kyoto University[96],[97]. The network activity has been tracked from the end of 2006 to December 2015. The dataset includes 24 features, 14 of them is based on the KDD Cup 99 feature vector and the last 10 features was added to better describe what happens on the network. The latter 10 features are described below: • Binary valued features that cover the observation of IDS alerts, Malicious connections and exploits of the network traffic. • Abnormal session feature state if the network traffic is benign or is a known/unknown attack. • Sanitized IP address, ports and session information was also captured. Obtain dataset here
62	crawdad	P	R	NET	N/A	crawdad is a compiled list of publicly available datasets of wireless protocols[98]. Obtain dataset here
63	ICS-pcap	P	N/A	NET	N/A	ICS-pcap: A repository of ICS/SCADA pcaps gathered from multiple sources[99]. Obtain dataset here
64	Common Crawl	R	R	NET	≈ 2000000 S	Common Crawl: A dataset of web content and metadata from 2000000 crawled webpages[100]. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
65	NSL-KDD	R	S	NET	N/A	NSL-KDD dataset: Is a subset of KDD99[90], [101] . Preproccsing that was performed on NSL-KDD: • Deduplication of training and testing set samples in NSL-KDD • The count of samples that is hard to classify in the NSL-kDD set, is inversely proportional to the percentage of hard samples in KDD. Obtain dataset here
66	ISCXTNT	R	R	NET	22GB	ISCX Tor-nonTor (ISCXTNT): Contains both normal and Tor traffic[102],[103]. The real Tor traffic was generated from network, email and filesharing protocols. And audio, and video streams from popular applications. The samples/traffic is assign a class label for what application or protocol they belong to. Obtain dataset here
67	ISCXVNV	R	R	NET	28GB	ISCX VPN-nonVPN (ISCXVNV) dataset: Similar to the ISCX Tor-nonTor dataset[104], [105]. It has non-VPN traffic and labelled VPN traffic from multiple applications and protocols. Obtain dataset here
68	ISCXIDS	R	H	NET	229712,8 MB	ISCX IDS (ISCXIDS): This dataset includes real traffic for IPv4, IPv6, UDP, TCP, ARP, DNS, ICMP, HTTP, SMTP, SSH, IMAP, POP3, and FTP generated by using agents that simulate users interacting with these protocols[106], [107]. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
69	YPFC	P	R	PASS	N/A	Yahoo Password Frequency Corpus(YPFC): A sanitized password frequency corpus that protect the privacy of the user accounts[108], [109] . 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. Obtain dataset here
70	VincentPassword	P	N/A	PASS	2000000 S / 20 MB	Password dataset with 2000000 samples[110]. Obtain dataset here
71	MBT08	P	R	RAM	N/A	Memory Buddies Traces(MBT): This dataset is built up of the memory contents from computers and servers. The dataset contains metadata of the underlying platform, metadata of live processes, and the hash values of the 4KB memory pages[111], [112], [113]. Obtain dataset here
72	NUSSC	P	R	SMS	87300 S	The NUS SMS Corpus (NUSSC) includes 55835 English and 31465 Chinese SMS messages[114], [115]. 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. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
73	WebbSC11	P	R	SPAM	≈ 350000 S/ 1GB compressed	The Webb Spam Corpus 2011(WebbSC11): A custom crawler was built to collect spam web pages[116], [117]. 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. Obtain dataset here
74	DITSSC	P	R	SPAM	1353 S	The samples from DIT SMS spam corpus (DITSSC) is a collection of reported SMS spam, by UK mobile users. The corpus is stored in a XML format. Unique entries in the collection was assured by performing case insensitive depublication[118],[119]. Obtain dataset here
75	TREC05-07	R	N/A	SPAM	3 C	Spam corpora from TREC. This server host 3 spam corpus from 2005-2007[120]. Obtain dataset here
76	Hewlett spam	P	R	SPAM	4601 S	A spam dataset created by Hewlett-Packard Labs[121]. The feature vector contains: • frequencies of words and characters, • the average length, max length and total count of ”uninterrupted sequences of capital letters” • class label Obtain dataset here
77	WEBSPAMUK07	P	R	SPAM	105896555 S	WEBSPAM-UK2007: A labeled spam dataset of 105896555 entries[122]. Obtain dataset here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
78	microblogPCU	P	R	SPAM	221579 S	microblogPCU Data Set: A labeled spam dataset[123]. Example features: • Microblog poster gender, username, user id • Number of followers • Number of reposts Obtain dataset here
79	TREC11	R	R	SPAM	16000000 S	TREC 2011 microblog dataset contains 16 million normal and spam twitter posts[124]. Obtain dataset here
80	SPAM/HAM	P	R	SPAM	273 MB/ 5780 S	Image Spam Dataset: contains images that is originated from a normal (HAM) email message or a spam message (SPAM)[125], [126]. This dataset can be used as a training and testing set in a SPAM or HAM classification problem. Obtain dataset here
81	phishtank	P	R	PHI	N/A	phishtank is a open community where the users can add, search and validate instances of network phsising [127]. Search repository here
82	millersmiles	P	R	PHI	N/A	millersmiles is a repository of phising samples and allows users to add to the collection or search though the repository[128]. Search repository here

. . . continued

I	Name	Acc	DT	Cat	Size	Description
83	PWDS15	P	R	PHI	2456 S	Phishing Websites Data: A collection of 2456 phishing websites[129], [130]. Example features in this dataset: • Long URL, Tiny url or abnormal url. • Special symbols in URL: '@', '//', and '-' • The count of dots in URL • Domain registration date • Protocol information and ports Obtain dataset here

Table 4: Datasets

1.5 contribution (how my paper differs from the two reviews)

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] 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`.

[5] 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`.

[6] 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`.

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

[8] 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`.

[9] 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/`.

[10] Brennan M, Greenstadt R. Practical Attacks Against Authorship Recognition Techniques. In: Proceedings of the 21st Innovative Applications 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`.

[11] 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`.

[12] 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`.

[13] 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`.

[14] 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`.

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

[16] 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.

[17] 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`.

[18] 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`.

[19] 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`.

[20] 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`.

[21] 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`.

[22] schonlau. Masquerading User Data;. Last accessed (DD/MM/YYYY) 01/10/2017. Available from: `http://www.schonlau.net/intrusion.html`.

[23] Wang k, Stolfo S. One-Class Training for Masquerade Detection. 2003 01;Last accessed (DD/MM/YYYY) 01/10/2017. Available from: `https://www.researchgate.net/publication/247054265_One-Class_Training_for_Masquerade_Detection`.

[24] netresec. Publicly available PCAP files; 2017. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `https://www.netresec.com/?page=PcapFiles`.

[25] malware-traffic analysis. A source for pcap files and malware samples...; 2017. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `http://malware-traffic-analysis.net/`.

[26] pcapr. Welcome to pcapr, where pcaps come alive.;. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `http://www.pcapr.net/home`.

[27] evilfingers. PCAP Repository; 2010. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `https://www.evilfingers.com/repository/index.php`.

[28] caida. CAIDA Data - Overview of Datasets, Monitors, and Reports; 2017. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `https://www.caida.org/data/overview/`.

[29] mining group C. Datasets;. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `http://csmining.org/index.php/data.html`.

[30] azsecure data. Get Data; •. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `http://www.azsecure-data.org/get-data.html`.

[31] Corpora D. Corpora;. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `https://digitalcorpora.org/corpora`.

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

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

[34] Kaggle. Hillary Clinton's Emails; 2016. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `https://www.kaggle.com/kaggle/hillary-clinton-emails`.

[35] Tatman R. Fraudulent E-mail Corpus CLAIR collection of "Nigerian" fraud emails; 2017. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `https://www.kaggle.com/rtatman/fraudulent-email-corpus`.

[36] Ruano-Ordas D. Corpus 200 Emails. 2015 3;Available from: `https://figshare.com/articles/Corpus_200_Emails/1326662`.

[37] Enrondata. Data;. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `https://enrondata.readthedocs.io/en/latest/references/data/`.

[38] 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`.

[39] 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/`.

[40] 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`.

[41] 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`.

[42] 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.

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

[44] Pozzolo AD, Caelen O, Johnson RA, Bontempi G. Calibrating Probability with Undersampling for Unbalanced Classification. In: IEEE Symposium Series on Computational Intelligence, SSCI 2015, Cape Town, South Africa, December 7-10, 2015; 2015. p. 159–166. Last accessed (DD/MM/YYYY) 26/09/2017. Available from: `https://www3.nd.edu/~dial/publications/dalpozzolo2015calibrating.pdf`.

[45] Andrea. Credit Card Fraud Detection: Anonymized credit card transactions labeled as fraudulent or genuine; 2016. Last accessed (DD/MM/YYYY) 26/09/2017. Available from: `https://www.kaggle.com/dalpozz/creditcardfraud`.

[46] K R S, Zareapoor M. FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining. 2014 09;2014:252797. Last accessed (DD/MM/YYYY) 01/10/2017. Available from: `https://www.researchgate.net/publication/266746615_FraudMiner_A_Novel_Credit_Card_Fraud_Detection_Model_Based_on_Frequent_Itemset_Mining`.

[47] purdue. Index of /data/credit_card;. Last accessed (DD/MM/YYYY) 01/10/2017. Available from: `https://www.cs.purdue.edu/commugrate/data/credit_card/`.

[48] cms. Dataset Downloads; 2017. Last accessed (DD/MM/YYYY) 01/10/2017. Available from: `https://www.cms.gov/OpenPayments/Explore-the-Data/Dataset-Downloads.html`.

[49] CMS. Open Payments Public Use Files: Methodology Overview & Data Dictionary; 2017. Last accessed (DD/MM/YYYY) 01/10/2017. Available from: `https://www.cms.gov/OpenPayments/Downloads/OpenPaymentsDataDictionary.pdf`.

[50] Lopez-Rojas E. Synthetic Financial Datasets For Fraud Detection - Synthetic datasets generated by the PaySim mobile money simulator; 2017. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `https://www.kaggle.com/ntnu-testimon/paysim1`.

[51] Lopez-Rojas EA. Applying Simulation to the Problem of Detecting Financial Fraud. , Department of Computer Science and Engineering; 2016.

[52] Lopez-Rojas EA, Axelsson S. BankSim: A Bank Payment Simulation for Fraud Detection Research; 2014. Available from: `https://www.researchgate.net/publication/265736405_BankSim_A_Bank_Payment_Simulation_for_Fraud_Detection_Research`.

[53] Lopez-Rojas EA, Axelsson S. Synthetic data from a financial payment system - Synthetic datasets generated by the BankSim payments simulator; 2017. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `https://www.kaggle.com/ntnu-testimon/banksim1`.

[54] 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`.

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

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

[57] 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`.

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

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

[60] 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`.

[61] 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`.

[62] 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/`.

[63] 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.

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

[65] 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.

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

[67] 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`.

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

[69] Avdiienko V, Kuznetsov K, Gorla A, Zeller A, Arzt S, Rasthofer S, et al. Mining Apps for Abnormal Usage of Sensitive Data. In: Proceedings of the 37th International Conference on Software Engineering. ICSE 2015; 2015. .

[70] Avdiienko V, Kuznetsov K, Gorla A, Zeller A. About MUDFLOW;. Last accessed (DD/MM/YYYY) 26/09/2017. Available from: `https://www.st.cs.uni-saarland.de/appmining/mudflow/`.

[71] Lab IR. Datasets; 2010. Last accessed (DD/MM/YYYY) 01/10/2017. Available from: `https://www.uvic.ca/engineering/ece/isot/datasets/index.php#section0-0`.

[72] ISOT. ISOT Dataset Overview;. Last accessed (DD/MM/YYYY) 01/10/2017. Available from: `https://www.uvic.ca/engineering/ece/isot/assets/docs/isot-datase.pdf`.

[73] Saad S, Traoré I, Ghorbani AA, Sayed B, Zhao D, Lu W, et al. Detecting P2P botnets through network behavior analysis and machine learning. In: PST. IEEE; 2011. p. 174–180. Available from: `http://ieeexplore.ieee.org/abstract/document/5971980/`.

[74] lirmm. Analyzing Web Traffic ECML/PKDD 2007 Discovery Challenge; 2007. Last accessed (DD/MM/YYYY) 01/10/2017. Available from: `http://www.lirmm.fr/pkdd2007-challenge/index.html#dataset`.

[75] isi csic es. HTTP DATASET CSIC 2010; 2012. Last accessed (DD/MM/YYYY) 01/10/2017. Available from: `http://www.isi.csic.es/dataset/`.

[76] contagiodump. Collection of Pcap files from malware analysis; 2015. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `http://contagiodump.blogspot.no/2013/04/collection-of-pcap-files-from-malware.html`.

[77] dropbox. PCAPS_TRAFFIC_PATTERNS fra DeepEnd Research (DeepEnd Research);. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `https://www.dropbox.com/sh/7fo4efxhpenexqp/AACmuri_l-LDiVDUDJ3hVLqPa?dl=0`.

[78] Dolan-Gavitt B. (Sys)Call Me Maybe: Exploring Malware Syscalls with PANDA; 2015. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `http://moyix.blogspot.no/search?q=dataset`.

[79] Garcia S. Dataset; 2015. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `https://stratosphereips.org/category/dataset.html`.

[80] Samani EBB, Jazi HH, Stakhanova N, Ghorbani AA. Towards effective feature selection in machine learning-based botnet detection approaches. In: IEEE Conference on Communications and Network Security, CNS 2014, San Francisco, CA, USA, October 29-31, 2014; 2014. p. 247–255. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6997492`.

[81] UNB. Botnet dataset;. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `http://www.unb.ca/cic/research/datasets/botnet.html`.

[82] Abdul Kadir AF, Stakhanova N, Ghorbani AA. In: Qiu M, Xu S, Yung M, Zhang H, editors. Android Botnets: What URLs are Telling Us. Cham: Springer International Publishing; 2015. p. 78–91. Available from: `https://doi.org/10.1007/978-3-319-25645-0_6`.

[83] UNB. Android Botnet dataset;. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `http://www.unb.ca/cic/research/datasets/android-botnet.html`.

[84] 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`.

[85] 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`.

[86] 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`.

[87] Romain F, Pierre B, Patrice A, Kensuke F. MAWILab Combining Diverse Anomaly Detectors for Automated Anomaly Labeling and Performance Benchmarking. In: ACM CoNEXT 10. Philadelphia PA;. .

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

[89] ”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.

[90] 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.

[91] 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.

[92] 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/.

[93] states military academy west point U. Data Sets; 2009. Last accessed (DD/MM/YYYY) 26/09/2017. Available from: Unitedstatesmilitaryacademywestpoint.

[94] Creech EITUCU Gideon. Developing a high-accuracy cross platform Host-Based Intrusion Detection System capable of reliably detecting zero-day attacks. Awarded by:University of New South Wales. Engineering and Information Technology; 2014. Last accessed (DD/MM/YYYY) 26/09/2017. Available from: http://unsworks.unsw.edu.au/fapi/datastream/unsworks:11913/SOURCE02?view=true.

[95] unsw. The ADFA Intrusion Detection Datasets; 2013. Last accessed (DD/MM/YYYY) 26/09/2017. Available from: https://www.unsw.adfa.edu.au/australian-centre-for-cyber-security/cybersecurity/ADFA-IDS-Datasets/.

[96] takakura. Traffic Data from Kyoto University’s Honeypots; 2015. Last accessed (DD/MM/YYYY) 01/10/2017. Available from: http://www.takakura.com/Kyoto_data/.

[97] SONG J, Takakura H, Okabe Y. Description of Kyoto University Benchmark Data;. Last accessed (DD/MM/YYYY) 01/10/2017. Available from: http://www.takakura.com/Kyoto_data/BenchmarkData-Description-v5.pdf.

[98] RAWDAD. All datasets and tools: sorted by name; 2017. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: https://crawdad.org/all-byname.html.

[99] automayt. A collection of ICS/SCADA PCAPs; 2016. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: https://github.com/automayt/ICS-pcap.

[100] amazon. Common Crawl on AWS;. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: https://aws.amazon.com/public-datasets/common-crawl/.

[101] UNB. NSL-KDD dataset;. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: http://www.unb.ca/cic/research/datasets/nsl.html.

[102] Lashkari AH, Gil GD, Mamun MSI, Ghorbani AA. Characterization of Tor Traffic using Time based Features. In: Proceedings of the 3rd International Conference on Information Systems Security and Privacy - Volume 1: ICISSP,. INSTICC. SciTePress; 2017. p. 253–262.

[103] UNB. Tor-nonTor dataset;. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: http://www.unb.ca/cic/research/datasets/tor.html.

[104] Draper-Gil G, Lashkari AH, Mamun MSI, Ghorbani AA. Characterization of Encrypted and VPN Traffic using Time-related Features. In: Proceedings of the 2nd International Conference on Information Systems Security and Privacy - Volume 1: ICISSP,. INSTICC. SciTePress; 2016. p. 407–414.

[105] UNB. VPN-nonVPN dataset;. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: http://www.unb.ca/cic/research/datasets/vpn.html.

[106] Shiravi A, Shiravi H, Tavallaee M, Ghorbani AA. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Computers & Security. 2012;31(3):357 – 374. Available from: http://www.sciencedirect.com/science/article/pii/S0167404811001672.

[107] UNB. Intrusion detection evaluation dataset;. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: http://www.unb.ca/cic/research/datasets/ids.html.

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

[109] 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.

[110] Granville V. Password and hijacked email dataset for you to test your data science skills; 2012. Last accessed (DD/MM/YYYY) 26/09/2017. Available from: http://www.datasciencecentral.com/forum/topics/password-dataset-for-you-to-test-your-data-science-skills.

[111] Wood T, Tarasuk-Levin G, Shenoy P, Desnoyers P, Cecchet E, Corner MD. Memory Buddies: Exploiting Page Sharing for Smart Colocation in Virtualized Data Centers. In: Proceedings of the 2009 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environ-

ments. VEE '09. New York, NY, USA: ACM; 2009. p. 31–40. Available from: `http://doi.acm.org/10.1145/1508293.1508299`.

[112] umass. Index of /traces/cpumem/memtraces; 2009. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `http://skuld.cs.umass.edu/traces/cpumem/memtraces/`.

[113] umass. readme.txt;. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: `http://skuld.cs.umass.edu/traces/cpumem/memtraces/readme.txt`.

[114] 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`.

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

[116] 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`.

[117] 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`.

[118] Delany SJ, Buckley M, Greene D. SMS spam filtering: Methods and data. Expert Systems with Applications. 2012;39(10):9899 – 9908. Available from: `http://www.sciencedirect.com/science/article/pii/S0957417412002977`.

[119] dublin institute of technology. DIT SMS Spam Dataset;. Last accessed (DD/MM/YYYY) 26/09/2017. Available from: `http://www.dit.ie/computing/research/resources/smsdata/`.

[120] NIST. Spam Track; 2017. Last accessed (DD/MM/YYYY) 26/09/2017. Available from: `http://trec.nist.gov/data/spam.html`.

[121] UCI. Spambase Data Set; 1999. Last accessed (DD/MM/YYYY) 26/09/2017. Available from: `http://archive.ics.uci.edu/ml/datasets/Spambase?ref=datanews.io`.

[122] WEBSPAM-UK2007. "Web Spam Collections"; 2007. Crawled by the Laboratory of Web Algorithmics, University of Milan, http://law.di.unimi.it/.Last accessed (DD/MM/YYYY) 26/09/2017. Available from: `http://chato.cl/webspam/datasets/uk2007/`.

[123] Jun Liu(liukeen '@' mail xjtu cn) MZJMYL Hao Chen(lechenhao '@' gmail com). microblogPCU Data Set;. MOEKLINNS Lab, Department of Computer Science ,Xi'an Jiaotong University, China. Last accessed

(DD/MM/YYYY) 26/09/2017. Available from: https://archive.ics.uci.edu/ml/datasets/microblogPCU.

[124] NIST. Tweets2011; 2014. Last accessed (DD/MM/YYYY) 26/09/2017. Available from: http://trec.nist.gov/data/tweets/.

[125] Dredze M, Gevaryahu R, Elias-Bachrach A. Learning Fast Classifiers for Image Spam.; 2007. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.8417&rep=rep1&type=pdf.

[126] jhu. Image Spam Dataset; 2007. Last accessed (DD/MM/YYYY) 02/10/2017. Available from: http://www.cs.jhu.edu/~mdredze/datasets/image_spam/.

[127] PhishTank. FAQ;. Last accessed (DD/MM/YYYY) 26/09/2017. Available from: http://www.phishtank.com/faq.php#whatisphishing.

[128] millersmiles. about us; 2017. Last accessed (DD/MM/YYYY) 26/09/2017. Available from: http://www.millersmiles.co.uk/aboutus.php.

[129] Mohammad R, McCluskey TL, Thabtah FA. Intelligent Rule based Phishing Websites Classification. IET Information Security. 2014 May;8(3):153–160. Available from: http://eprints.hud.ac.uk/id/eprint/17994/.

[130] UCI. Phishing Websites Data Set;. Last accessed (DD/MM/YYYY) 26/09/2017. Available from: http://archive.ics.uci.edu/ml/datasets/Phishing+Websites#.