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. In the list below is examples of relevant datasets:

- forensics images,
- files corpora,
- RAM dumps,
- network files,
- malware,
- email,
- SMS,
- phishing,
- spam,
- authorship classification files,
- financial data/ fraud,
- forgery corpus,
- language files etc.

Decisions was made to limit the scope of the data collection, by excluding biometric datasets (e.g. images of fingerprints).

During the collection phase of the literature review, two related reviews was identified.[source 1(old), source 2(new) (come back too)] ... I expand upon the existing work ....

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.

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

[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

Table 1: Search summary

1.4 Search summary - datasets:

Table 2 is a summary of the identified datasets. An entry in this table is explained in the list below:

- Column I = Numbered Item
- 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 MAL=Malware, NET=Network etc
- Column Size= Size is either given in samples, GigaBytes (comrpressed/uncompressed) or unknown
- 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.

I	Acc	DT	Cat	Size	Description
1	P	R	MAL	0.05	
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					

Table 2: Datasets