Technical report on

An Empirical Investigation of Paedophile Keywords in eDonkey P2P Network

Measurement and Analysis of P2P Activity Against Paedophile Content project
http://antipaedo.lip6.fr

dr. Vasja Vehovar, dr. Aleš, Žiberna, dr. Matej Kovačič, dr. Andrej Mrvar, May Doušak
University of Ljubljana, Faculty of Social Sciences, Centre for Methodology and informatics

SUMMARY

We used the eDonkey datasets provided by MAPAP project and performed research in two direction: (1) exploratory analysis of structure and regularities related to paedophile keywords, files and IPs and (2) discovering of new paedophile keywords.

1. Exploratory analysis

Keywords. We started with initial set of 21 contaminated (i.e. paedophile-related) words identified arbitrary by non-expert. The selection was thus based on common sense imagination, so the keywords were very much obviously related to paedophile content. Next, fourteen (14) additional words were added via frequent appearance in search strings together with original 21 keywords. The set of 35 keywords also included half of keywords used by French police. On the other hand, the match with INHOPE keyword lists from UK (IWF) and US (NCMEC), which contained 482 and 393 keywords respectively , was surprisingly low - only 6 and 7 keywords overlaped.

Queries. In total 35 million queries were performed with 120,000 different words in search strings of our datasets; majority of keywords had the frequency above 100. Out of them 11,000 keywords were at least once in a string with contaminated keywords.

Search IPs. Roughly around 0.2% (21,000) of all IPs (11 mio) made queries using contaminated keywords – we call them contaminated search IPs. Only 3300 IPs used more than one keyword and 70 more than five, maximum was 14 keywords.

Files. Roughly around 0.3% (around 20,000) of all files (around 7 mio) were searched for with contaminated keywords – we call them contaminated files. For 272 files two and for 19 files three different keywords were used. Of course, these files were also searched by other keywords; a median was 50 keywords. The contaminated keywords presented only 5% share among all keywords used to find the contaminated files. However, for 2,400 files the contaminated keywords present more than 30% of all keywords used to hit these contaminated files.

Supply IPs. First, we should expose that in our data there were only 1 million IPs which hosted at least one file, while 11 millions IPs performed at least one search. The overlap was only 3000 files, due to the fact that we used data from only one of many eDonkey servers. Al searches were thus fully recorded, while the supply files might come from other servers. Another explanation could arise from dynamic IP.

Around 3% (365 000) of all IPs (12 mio) had at least one contaminated file – we label them contaminated supply IPs. However, if we take into account only IPs which hosted files (i.e. 1 mio IPs, with median of 19 files per IP and maximum 5,366), the percentage rose to 36% (365,000 IPs, maximum 36 files per IP). Among them 18,000 IPs had 5 or more contaminated files and 3,000 have more than 50% of files being contaminated. However, only 11 IPs had contaminated files with more than 50% of contaminated keywords among all keywords searching them. On average, each contaminated file was hosted by 18 supply IPs. Only 334 of the all files were hosted by contaminated suppy IPs, which also used contaminated keywords in search quires. Each of these files was hosted by only 1 contaminated IP. Only 3 IPs were both searching with contaminated keywords and hosting the contaminated files.

Networks. Disposing with contaminated files, keywords, search IPs, supply IP2 we used social network software ***Pajek*** to discover hidden network structure. We did discover numerous and various structures, however all networks were very small.

Conclusions:

- Likelihood for file or search IP to relate with paedophile keywords is around 0.2.
- Treating a file or search IP as contaminated basing on only one contaminated keyword seem to be too broad. Sharper restrictions (i.e. more contaminated keywords/searches needed to declare contamination) would shrink the above estimates by factor 10, what can be treated as a hard core, where little doubt about paedophile nature remains. While the estimated is thus in 0.02% - 0.2%, further iterations of our approach (i.e. expanding/refining the keyword list) would stabilize and narrow this interval to the upper bound and might even surpass it. In total, however, the paedophile appearance of this size is not negligible at all, despite certain technical and methodological limitations.
- The contaminated supply IPs seem rarely be also the contaminated search IPs. In large part, this is due to the fact that we observed all active search IPs from one eDonkey server only, while supply IPs may come via all other eDonkey servers, where searches were not observed. Dynamic IPs may also contribute to this.
- We did not discover any well-articulated network of contaminated elements (keywords-files-IPs) of substantial size, but only relatively small sub-networks.
- We targeted here a very general paedophile users and suppliers, who do not work on this content systematically and are thus not the organized professionals.

2) Additional keyword analysis

We have analyzed searches made by contaminated IPs, where share of contaminated words presented majority of their keywords. Based on that, we proposed 68 new contaminated keywords. Considerable part of them could be immediately confirmed with a simple web search (e.g. madebyarkh).

We also studied the searches that lead to contaminated files and analyzed the appearance of other potentially contaminated keywords. We thus obtained 58 new potentially paedophilic words; many of them being directly recognized by simple web search (e.g. reelkiddymov) or being already included into the police lists.

The overlap among the two additional sets of keywords was extremely small.

This approach thus proved to be very fruitful in finding additional keywords related to the paedophile content in P2P networks. Iterating this approach would further increase the number of contaminated elements (keywords, files, search IPs and supply IPs) and would also converge to a stabile sizes for all those sets of elements.

INTRODUCTION

A. The core methodology

We analyzed the dataset provided by CNSR at their home page on April 2008, i.e. One week measurement on an eDonkey server, followed by corresponding description Technical description of measurements on eDonkey servers and using the corresponding labelling/decoding for the frequently used keywords.

The following potential approaches were considered for our analysis:

1) **Data mining:** We may use a data miming tool, e.g. globally recognized tool "Text-Garden". It is a Data Mining Software Tool designed for text document analysis. The search queries are effectively short documents, and the OntoGen semi-automatic and data-driven ontology construction tool allows to construct an ontology of queries with an efficient user interface. Users, files and keywords can be modelled as individual terms and enable to discover indirect links between them. With this tool we can effectively identify types of keywords and prepare visualizations of the content.

2) **Topic modelling** is a recent development from the legacy of latent semantic analysis (LSA). Topic models are an effective way to capture correlations between keywords, thus forming topics. In that respect, they may supplement networks: networks are effective illustrations of relationships between a small numbers of keywords. For a large number of correlated keywords, it helps to replace all of them with the notion of a topic. A large number of correlated keywords are thus reduced to a smaller number of independent or complementary topics, but each of those topics can be examined in more detail to obtain information about the keywords forming it. Topic models have recently been used in combination with social network analysis (e.g. http://cosco.hiit.fi/Articles/wi04chat.pdf)

3) **Social network analysis**. Here we may use Pajek tool, developed at the Centre for Methodology and Informatics, Faculty of Social Sciences, University of Ljubljana, which is the leading software for analyzing large social networks. As for now, it can handle up to 10 million nodes.

Due to the strong relational nature of our data we decided to predominantly use social network approach.

B. Conceptual layout

The basic idea of our research is to study the structure of paedophile elements in eDonkey P2P network and also to identify new keywords, which may be used in searches that lead to illegal content. We assume that besides obvious keywords (e.g. "young girls") there also exist some other "contaminated" keywords (related to paedophilic content but known only to insiders), which can further lead to "contaminated" users and to "contaminated" data-files. We focus on these indirect information linkages. After creating the initial common sense list of obvious keywords we proceed in steps as follows.

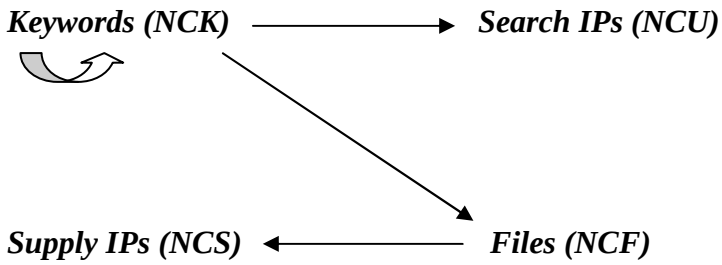
1) **FIRST STEP:** In the first step we define sets of contaminated elements (files, keywords, supply IPs and search IPs) starting from contaminated keywords.

a) The users who used an obvious contaminated keyword (e.g. "young girls") may also use other typical keywords in their own search strings we are not yet aware of. So the network of (additional) contaminated keywords (NCK) used within the same search strings together with initial (common sense) keywords will be created.

b) In addition we also study the network of users (search IPs) who used the contaminated keywords, so we obtain the network of contaminated users (NCU).

c) Further, we study the network of contaminated files (NCF) accessed via contaminated keywords (i.e. NCK).

d) The suppliers of contaminated files (i.e. supply IPs) can be identified as network of contaminated suppliers NCS.



2) **SECOND STEP:** Based on initial sets of contaminated files from previous step, the sets of keywords search IPs and supply IPs (i.e. NCK, NCF, NCU and NCS) are further expanded.

a) With NCF we can indirectly observe other keywords that had also led to the same contaminated files, what can expand the initial set of contaminated keywords NCK.

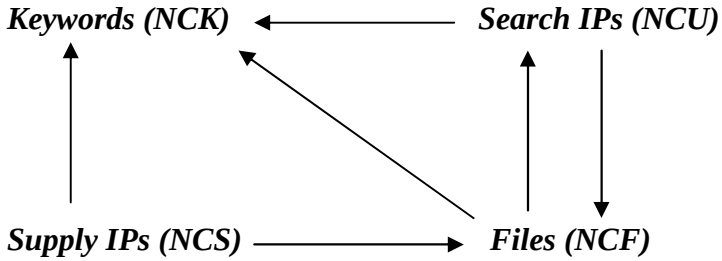
b) Similar expansion of NCK can be obtained by observing other keywords used by contaminated search IPs (i.e. NCU), that is, frequent keywords used by already defined contaminated initial set of contaminated search IPs (NCU).

c) The initial set of contaminated users (i.e. search IPs) also defines additional files (besides those directly targeted with contaminated keywords), which are also searched by these NCU and are thus potentially contaminated. Initial set of NCU leads to expanded NCF.

d) Similarly, the initial set of contaminated files directs to other users searching for these same contaminated files (regardless of the contamination level of corresponding keywords). The initial NCF thus leads to expanded NCU.

e) Finally, the initial set of suppliers (i.e. NCS) directs to potentially contaminated keywords that these supplies IPs are using in their own searches. The initial NCS thus expands NCK.

f) The contaminated supply IPs (i.e. NCS) also have other files that they are hosting, what gives those files increased likelihood of being contaminated. So the initial NCS expands NCF.



3) ITERATIONS

We can continuously repeat the two steps above, one after another. The looping of networks thus generates extended NCK, NCF, NCU and NCS and can be used to interatively generate second-level, third-level etc networks of keywords, files, users, supplier (i.e. NCK, NCF, NCU and NCS), until we reach the convergence and the desired stability.

Of course, if we want to perform iterations that would converge, each potentially contaminated element (keyword, file, user IP, supply IP) need to be assigned a propensity score (probability/odds of being contaminated) at each step, calculated as a compound measure of its value from the previous step combined with the scores from neighbouring/related elements obtained in the previous step.

The starting values in the FIRST step can be relatively arbitrary assigned, according to some very simple but reasonable rule, e.g. the propensity score for a file being contaminated is proportional to the number of hits the file received by searches using contaminated keywords etc.

Nevertheless, due to the complexity of these algorithms and due to limited resources, we only performed here the entire FIRST step, i.e. actions (a), (b), (c) and (d), while from the SECOND step we performed in our analysis all the actions that directly expand the keyword lists, i.e. actions (a), (b) and (e).

C. Technical issues

In the first stage of our research the technical problems were dealt due to large dataset. The solutions to the optimal organization of the data were sought and proper scripts were written in Python to transform the existing data format into optimal form suitable for network and data-mining analysis. This task was relatively demanding – due to extremely large datasets - and required much more work than anticipated.

The right approaches were sought to address this very specific problem, because wrong selection may lead to no or little results with substantive loss of resources.

Due to extremely large datasets we limit our research only to ***half*** of the total data, so we shrunk the observed time where the logs from eDonkey server were observed by half. This enabled us to fully performed network analysis.

In the total MAPAP eDonkey dataset there were thus approximately twice more elements (IPs, files, searches) compared to the size of these datasets in our analysis, where we were dealing in total with around ***12 millions of IPs***, ***35 millions of search queries*** and ***7 millions of files***.

ANALYSIS

1. Network analysis of initial keywords

We analyze the networks of the search query terms used in the eDonkey network (obtained from anonymized_strings.gz). Our intent is to identify “common” words that are most frequently used in search queries with the “paedophile” words (i.e. “contaminated” words) and look at the connections among them. Then we might repeat the search for the “neighbours” (the words that appear in the queries with these words) of this new selection of words that would include the original “paedophile” selected words and the words that are most heavily connected to them. With this analysis we will try to find “contaminated” words that are often used in queries related to paedophilia. The goal of this task is to find out whether there are some “secret” words, that paedophiles use to find illegal content on the P2P networks, but is by them ordinary, everyday words or they usually search for the illegal content with “direct” search terms.

word	*frequency*
abuse	3488
abused	733
boychild	107
childlover	707
childporn	188
kidnap	430
kidnapped	1292
kidnapping	456
necrofilia	119
pedo	11413
pedofilia	2318
pedofilo	119
pedoland	104
paedophile	117
pedophilia	103
pedos	134
youngmodels	219
youngporn	351
zofilia	117
zoofilia	3005
zoophilia	270

Table 1: Selected words used in eDonkey search queries.

Here we start with preliminary analysis of 21 “contaminated” keywords to test the performance and behaviour of the computer process. For the analysis we first had to

de-anonimize search queries. Since we only have a list of words which appear in search queries more than 100 times, we generated de-anonimized search queries list. Words we can not identify, we present as a number (for instance v1594). From the search query terms used, we first selected 21 words which are more or less clearly connected to paedophilia. These words with frequency are presented in Table 1.

Quick analysis shows that these words appear as search terms 25.790-times. To find out which words are connected with them, we selected all words that appeared at lest once in the same query as one (or more) of the original 21 selected words. That selection was a basis for building the network.

That was the basis to create a network of words that included all connections among all words in this extended selection of words. A tie among words is created, if the two words appeared together in at lest one search query. The number of such words (including the "paedophilic" ones) is 11250, while the number of ties among them is 6329873, of which 6960 are loops. The value of the tie between two words represents the number of times the two words appeared together in the same search query. The 10 highest tie values are presented in Table 2.

rank	*tie value*	*tie (the endpoints)*
1	1135274	the-mp3
2	946968	the-of
3	663139	la-de
4	512184	de-mp3
5	430627	mp3-a
6	426764	of-mp3
7	417148	la-mp3
8	394320	the-the
9	344140	mp3-i
10	343260	in-the

Table 2: Ten ties with the highest tie values.

As this value heavily depends on the frequency of the two words in general, we generated the weighted version of this network. The weight is computed using a so-called "jaccard" coefficient or more precisely using the following formula:

$$w_{ij} = \frac{t_{ij}}{n_i + n_j - t_{ij}}$$

where t_{ij} is the number of times words i and j appeared together and n_i is their frequency of word i.

Building of the network of words and network analysis is very computer and time intensive process. To optimize the analysis we removed all ties with tie values lower than 0,01. After that, only 24188 ties remain of which 942 are loops. For network analysis Pajek software was used and input data has been prepared with Python scripts.

Figure 1 shows connections among initial 21 "paedophilic" words with tie weight above 0,01 in the jaccard network. Quick analysis show strong ties among words "pedo", "pedofilia" (and all its variants) and "childlover", "childporn", "zoofilia"and

"abuse". That means people searching for pedophilia also search for "zoofilia", "abuse", "childlover", etc. in the same query.

In Figure 2 we show connections among initial 21 words and those connected to them by tie with weight of at least 0,01. To make the presentation clearer we removed all ties with low tie values (lower than 0,01) also among the selected words. Quick view of the layout shows that words connected with the initial "paedophilic" words are mainly "contaminated" (for example: "incest", "lolita", "brutalviolence", "gag", "mafiasex", "kinderficker", etc.).

In Figure 3 we also present a layout showing all connections (even those with weights below 0,01) among the selected words. In Figure 4 we present a layout where all ties between "non-paedophilic" words were removed.

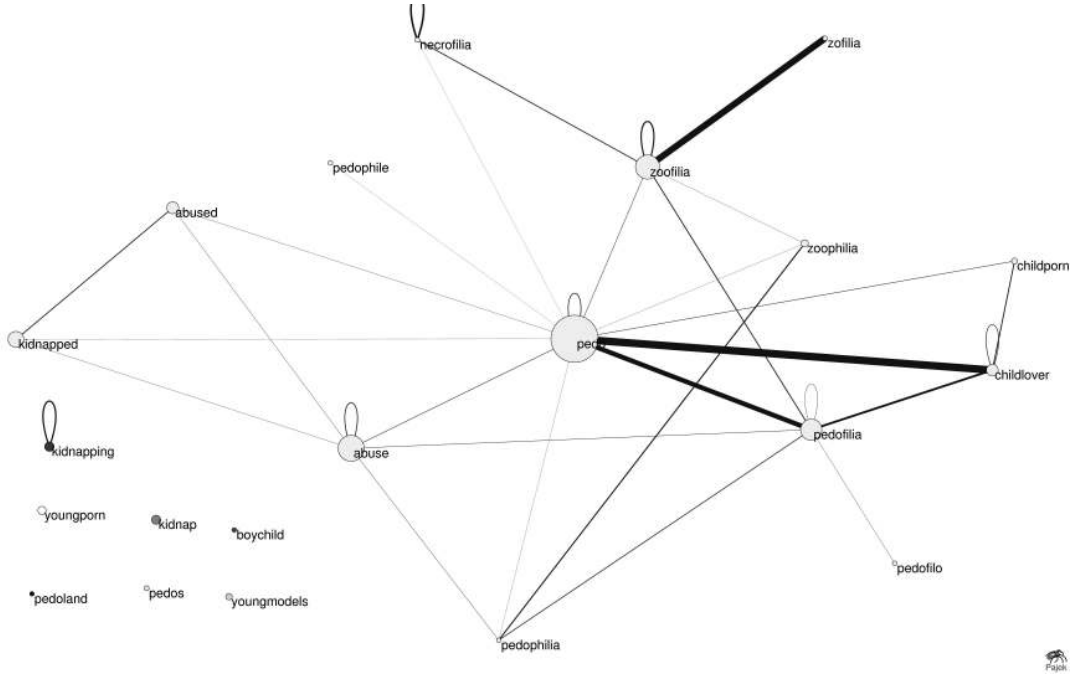


Figure 1: Layout showing connections among 21 "paedophilic" words with tie weight above 0,01 in the jaccard network. Loops on a vertices means that the same word appeared in search query two or more times. The value of the largest weight is 0,018.

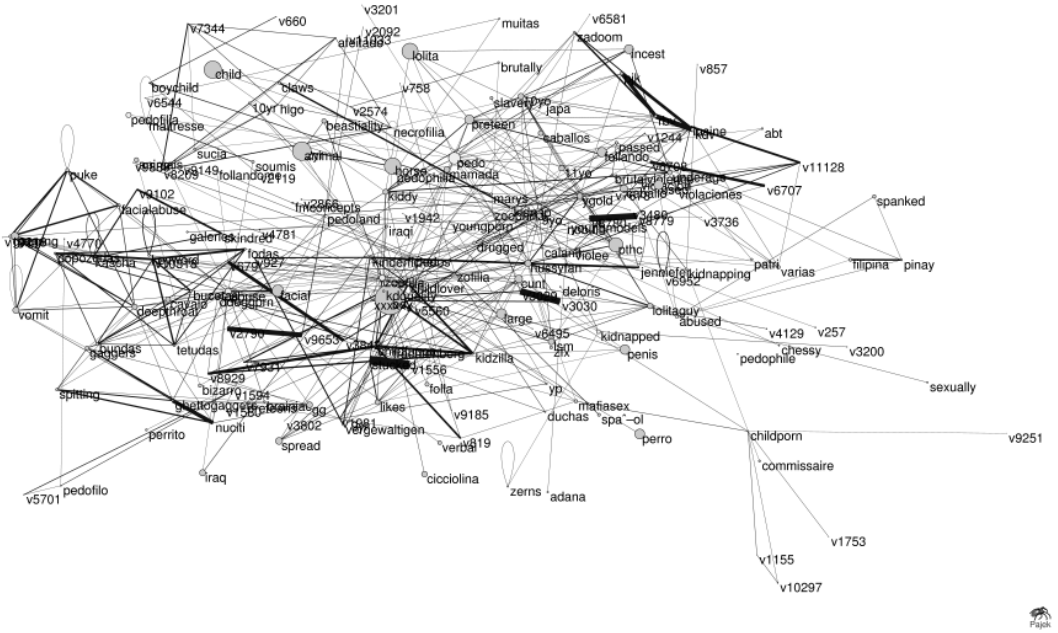


Figure 2: Layout showing connections with weight of at least 0,01 among 21 words and those connected to them by tie with weight of at least 0,01. The value of the largest weight is 0,517.

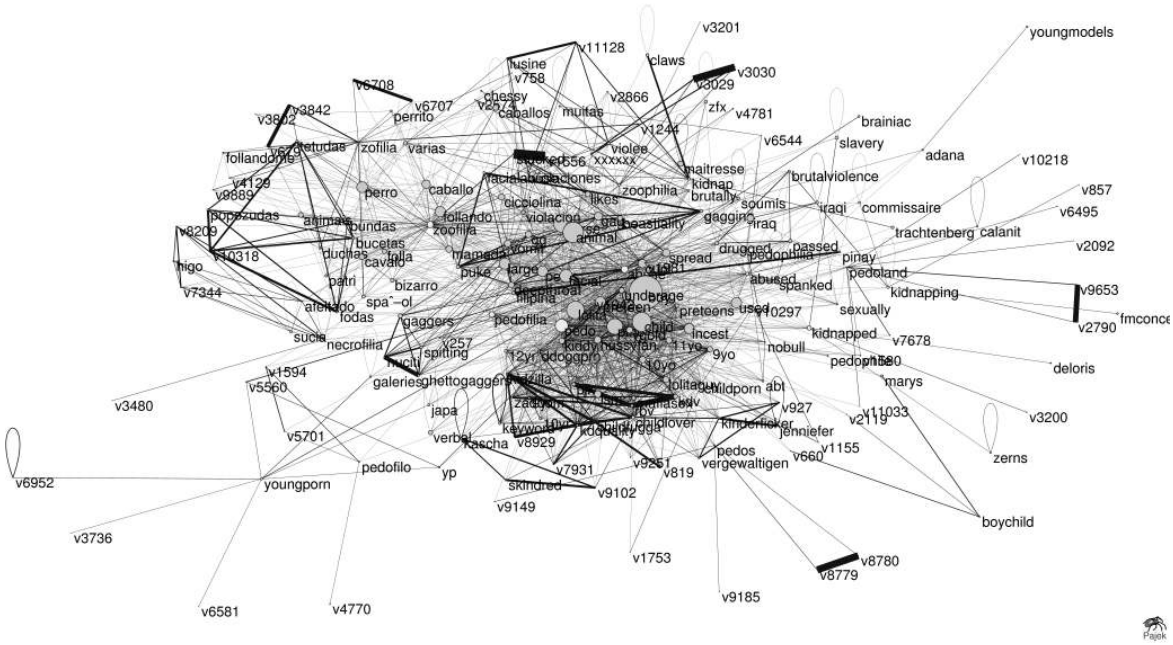


Figure 3: All connections among 21 words and those connected to them by tie with weight of at least 0,01. The value of the largest weight is 0,517.

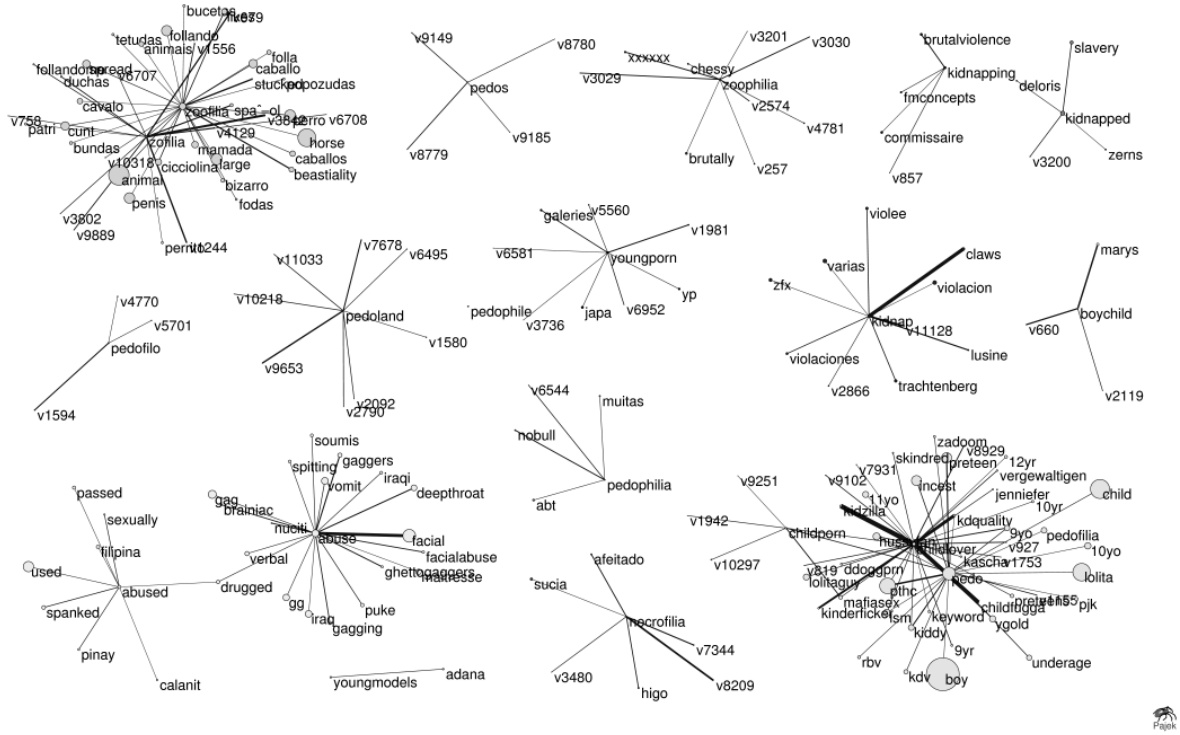


Figure 4: Layout of 21 words and those connected to them by tie with weight of at least 0,01 with connections only among the original 21 words. The value of the largest weight is 0,123.

2. Additional keywords

With this basic analysis we identified several other words that could be potentially used to find paedophilic material. These are:

1. QWERTY is a secret code word used by paedophiles and porn junkies. It is added to the end of file names as a method to return more porn results when using file sharing programs such as WINMX. Also used to disguise illegal child porn files. If you can`t find the porno your looking for, just try searching for "QWERTY." (http://www.urbandictionary.com/define.php?term=qwerty)
2. childfugga (meaning is unknown, however results obtained using this keyword in google search engine indicate that word could be connected to child pornography)
3. kinderficker - Child Molester (in German) (http://www.urbandictionary.com/define.php?term=kinderficker)
4. kidzilla - Acronym for underage porn, used in p2p searches. Simular to kidzilla: PTCH (preteen hardcore), Babyj, hussyfan, lolita. (http://www.urbandictionary.com/define.php?term=kidzilla)
5. kiddie (http://www.urbandictionary.com/define.php?term=kiddie):
    1. Anyone on the Internet under the age of 13 that acts like they run the place. E.g.:GameFAQs.com is full of kiddies, each stupider then the last.
    2. kiddie - What a pedo wants
6. ddoggprn (meaning is unknown, however results obtained using this keyword in google search engine indicate that word could be connected to child pornography)
7. kdquality - kiddy quality porn. "spamless" cp search (http://www.urbandictionary.com/define.php?term=kdquality)
8. ptsc - Pre teen soft core (Child sexual abuse pornography terms.xls)
9. hussyfan - hussyfan is a keyword used in p2p systems by children who want to find pictures of people their age instead of looking at adult porn. hussyfan - hussyfan is one of many p2p acronyms for underage pornography like PTHC or R@ygold (http://www.urbandictionary.com/define.php?term=hussyfan)
10. pthc - stands for Pre Teen Hard Core,used in most p2p networks to download (http://www.urbandictionary.com/define.php?term=pthc)
11. babyshivid – Aslo connected to child pornography (http://www.urbandictionary.com/define.php?term=babyshivid)
12. ygold – This is the second part of the word r@ygold and is present,a s non-alphanumeric characters were removed/used to split words – r@ygold: Actually NOT a real person; R@ygold is simply a codename used by paedophiles so that they can easily locate each other`s media. R@ygold is a keyword added to image and video files with illegal pornographic content, so that those dealing in child porn can locate and share files over P2P networks. (http://www.urbandictionary.com/define.php?term=r%40ygold)
13. lolitaguy - (meaning is unknown, however results obtained using this keyword in google search engine indicate that word could be connected to child pornography)

14. nymphets - (meaning is unknown, however results obtained using this keyword in google search engine indicate that word could be connected to child pornography)

Our results also indicate that especially IDs 8929 and 5927 and perhaps also IDs 7931, 7107 and 819 could be used to find child pornography. As they are not observed at least 100 times their meaning is unknown.

We are also wondering, what the following words could mean, as they are highly correlated with contaminated keywords:

1. lso
2. lsbar
3. 001a

3. Identification of contaminated files and IPs

3.1 Final selection of keywords for identification of paedophilic files and IPs

In addition to the words selected from the start and those selected based on the network analysis of the co-occurrences of words in queries we also considered the lists of paedophilic words received from The National Center for Missing & Exploited Children (NCMEC) from US (http://www.missingkids.com/) and Internet Watch Foundation (IWF) from UK (http://www.iwf.org.uk/) at the end of 2008. Unfortunately, those list words did not prove very helpful. The reason for that is that very little of those words were used often (more than 100 times) in queries on the eDonkey server. In Table 3 there are some statistics about words from these lists in comparison to the words for which we have the IDs used in the eDonkey messages.

	US (NCMEC)	UK (IWF)
Number of words in the lists	393	429
Number of "simple words" - words that do not contain non alphanumeric characters (including spaces)	172	192
Number of words for which we can get IDs	6	7

Table 3: Some statistics on words in UK and US lists of paedophile words.

The matched words were as follows:

- UK: hussyfan, jailbait, lolitas, lolitasex, lolly, pedoland, ptsc
- US: hussyfan, lolitasex, lolly, pedoland, 5, ptsc

In a final list of potentially paedophilic wordse we excluded a few words for we believed that could be frequently used also in other contexts. We ended with the list of words presented in Table 4. The jaccard similarity network among these words (based on queries) is presented in Figure 5. This list may still contain too many words that could be used in non-paedophile context. If we want to count single hits as indications, a more conservative list might be more appropriate.

For each IP/file we checked how many keywords were used in searches (for IPs – searches made by IPs, for files – searches using which files were found), how many of them were potentially paedophilic (based on the list described above) and what is the ratio of paedophilic words to all words used in searches. This was repeated using only unique words. When all words are used (not only unique), then if the same word appears twice, it is also counted twice. These statistics are used as they can show in a certain way how often paedophilic words are used (especially compared to other words).

We used the list of words in Table 4 to identify IPs that might search for paedophilic content (based on the keywords used) and files that might contain paedophilic content.

ID	String	freq
2065	lolitaguy	2647
2599	qwerty	1388
14239	ddoggprn	329
15999	pedo	11413
16587	hussyfan	3502
21847	pthc	17153
22211	lolitasex	633
22557	lolitas	5022
26029	ygold	2771
28846	ptsc	3171
31894	nymphets	653
37439	childlover	707
43019	babyshivid	699
53842	zoophilia	270
67057	pedophilia	103
70781	kinderficker	206
75499	kdquality	376
81306	paedophile	117
91985	pedofilia	2318
112145	kidzilla	254
126905	kiddie	426
134000	pedofilo	119
185184	pedoland	104
185684	pedos	134
201166	childporn	188
361730	childfugga	113

Table 4: IDs, words and their frequency that were used for identification of potentially paedophilic content

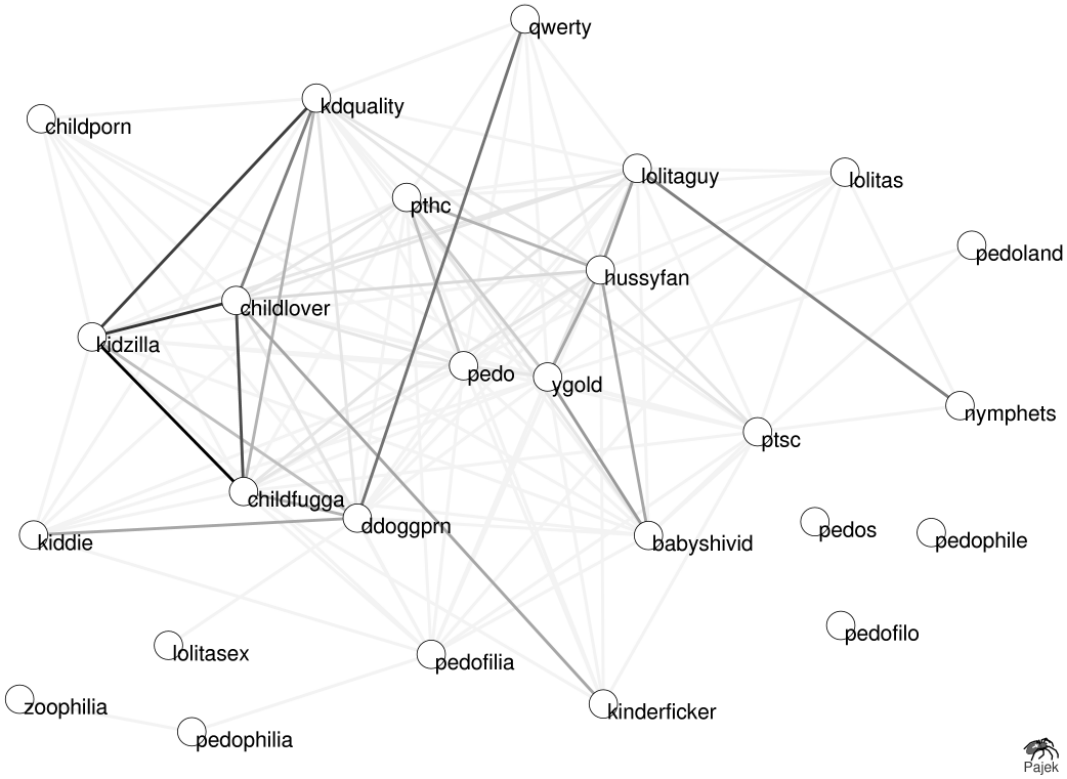


Figure 5: Layout showing connections among 26 "paedophilic" words with tie weight above 0,01 in the jaccard network. The value of the largest weight is 0.1505.

3.2 Files, keywords, searches and IPs: The specifics of the data

We analyzed the data available on http://content.lip6.fr/latapy/edonkey/weeks/weeks/ from 01/01/17 (week/day/hour) to 04/01/16 (both endpoints included) (three weeks). These are not dates, but consecutive numbers from the above web page. We do not know the exact dates when the data were collected, however based on Aidouni, Latapy and Magnien (2008b) the collection started in 2007.

We analyzed the data from OP GLOBSEARCHREQ, OP GLOBSEARCHRES and OP_GLOBFOUNDSOURCES (and similar) messages (see Aidouni, Latapy and Magnien (2008b) for content of these messages). Form these xml data we extracted (and used) the following informations:

1. Which words were used by each IP to find files
2. Words words were used to find each file (by IPs) – or with which words was this file "hit" or found
3. Which files were hosted by each IP and which IPs hosted each file

The data so collected include data on:

- 12 270 786 IPs
- 8 991 268 files
- 119 869 words with known IDs (or IDs with unknown words) – not necessaryly all present in the data
- 2 229 659 unique IDs present in the IP searches.

The IPs in particular can take on two roles – searchers and hosts. However, it seems that most of the IPs do not perform both roles, as can be seen in Table 5. Obviously, there is no IPs that would not take on at least one role, as in such case they are not included in the data. However we can see that most of the searchers do not host files and most of the hosts do not search for files.

		hosts	
		No	Yes
searches	No	0	1020036
	Yes	11248000	2750

Table 5: The roles that the IPs play in the eDonkey network.

In considerable part this can be explained by the fact that we obtained eDonkey data from only one eDonkey server (out of more than 50 at that time). We thus have log files for all searches the users (IPs) performed at this server and also all supply actions that this users provide. However, a lot of files were supplied to the users of this eDonkey server via other eDonkey servers, where we do not track their search activities. This is no doubt a considerable deficiency of these data, as we do not dispose with all eDonkey network activities of the users that appear in our datasets.

In part, the dynamic IPs may also contribute to the problem. There, user gets new/different IP number at each session or each day. We encounter this at various individual modem type of internet access and also at internal dynamically allocated IP numbers in large organisation's computer networks.

3.3 Files, keywords, searches and IPs: Basic demographics

We checked which IPs have potentially paedophilic files and which files are hosted (in possession) by potentially paedophilic IPs. For each IP we extracted the following data:

a) the number of files hosted (in possession),

b) the number of potentially paedophile (at least one potentially paedophilic keyword) files that an IP has,

c) the share of potentially paedophile files that an IP has among all files that an IP has d) the average number of unique potentially paedophile words in potentially paedophilic files that an IP has

e) the average share of unique potentially paedophile words in potentially paedophile files that an IP has.

Similar data was also gathered for files.

The information on which IPs have a given file was extracted from the server responses. Obviously, we only know which IPs have files which were queried by the users. The information on which files a given IP has was then obtained by transforming these data.

Using this procedure we identified only 2 IPs and 4 files that were “double positive”, meaning that they were identified as potentially paedophilic based on both criteria:

- the search words used by them (IPs) or to find them (files) and
- connections to files (IPs) or IPs (files) there were also identified as paedophilic

However, if we inspect the Tables 6 and 7 on the next page, we can see that even these IPs and 6 files can probably not seriously termed “paedophilic”.

Therefore, we have also looked at those files/IPs that were identified as potentially paedophilic based only on one criteria.

3.4 Identification of contaminated files

In the data analyzed, there were 8991268 files, of which we had data about keywords used to find them for 6819038 files. Of these 6819038 files for which we have data on keywords used to find them, the potentially paedophilic words were used in searches for 20519 of them. In Figures 6 and 7 we can see the distribution of number of (unique/different) words used in searches for files for which we have data on keywords used. We can see that a large number of words can be used to search for some files, for some files even 4007 different words were used in searches.

The distribution of the number of potentially paedophilic words is shown in Figure 8. For most of these files (20228) only one unique word was used to find them (and for most of them (18965) this one word was used only once). Only for 272 files two words were used and only for 19 files 3 different words were used.

fileID	Number of keywords used in searches	Number of potentially paedophilic keywords used	Share of potentially paedophilic keywords used	Number of unique keywords used	Number of unique potentially paedophilic keywords	Share of unique potentially paedophilic keywords	Number of IPs that have this file	Average number of potentially paedophilic keywords in potentially paedophilic IPs	Average share of potentially paedophilic keywords in potentially paedophilic IPs		
62978	395	1	0.002532	198	1	0.005051	470	1	0.002128	4.000000	0.363636
0000260291	229	1	0.004367	88	1	0.011364	73	1	0.013699	1.000000	0.055556
0000354761	261	1	0.003831	131	1	0.007634	302	1	0.003311	2.000000	0.181818
0000653555	60	1	0.016667	43	1	0.023256	91	1	0.010989	1.000000	0.055556
0005260650	39	1	0.025641	27	1	0.037037	53	1	0.018868	1.000000	0.055556
0010275505	184	1	0.005435	109	1	0.009174	127	1	0.007874	4.000000	0.363636

Table 6: The list of "double positive" files

IP ID	Number of keywords used in searches	Number of potentially paedophilic keywords used	Share of potentially paedophilic keywords used	Number of unique keywords used	Number of unique potentially paedophilic keywords	Share of unique potentially paedophilic keywords used	Number of files that this IP has	Number of potentially paedophilic files that this IP has	Share of potentially paedophilic files	Average number of potentially paedophilic keywords in potentially paedophilic files	Average share of potentially paedophilic keywords in potentially paedophilic files
0010793709	18	1	0.055556	17	1	0.058824	40	3	0.075000	1.000000	0.015558
0014348830	11	4	0.363636	3	1	0.333333	70	2	0.028571	1.000000	0.003984
0025104376	11	2	0.181818	5	1	0.200000	93	1	0.010753	1.000000	0.003831

Table 7: The list of "double positive" IPs

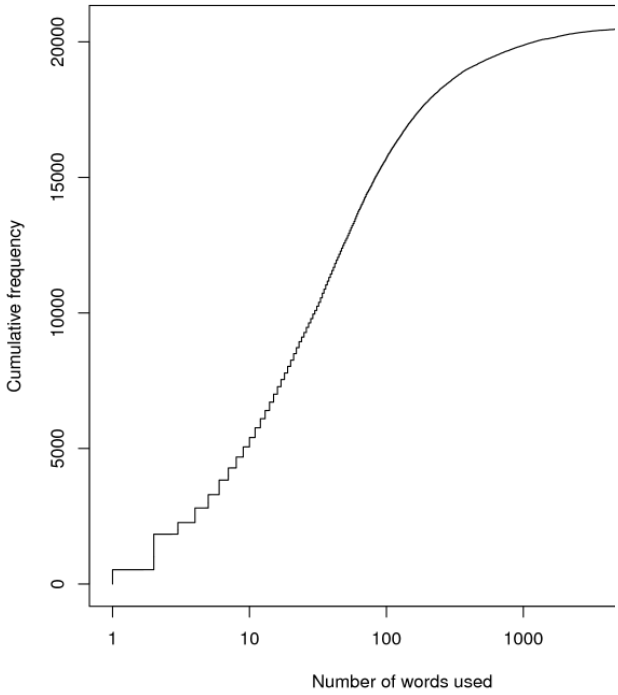


Figure 6: The number of words used in searches for a given file

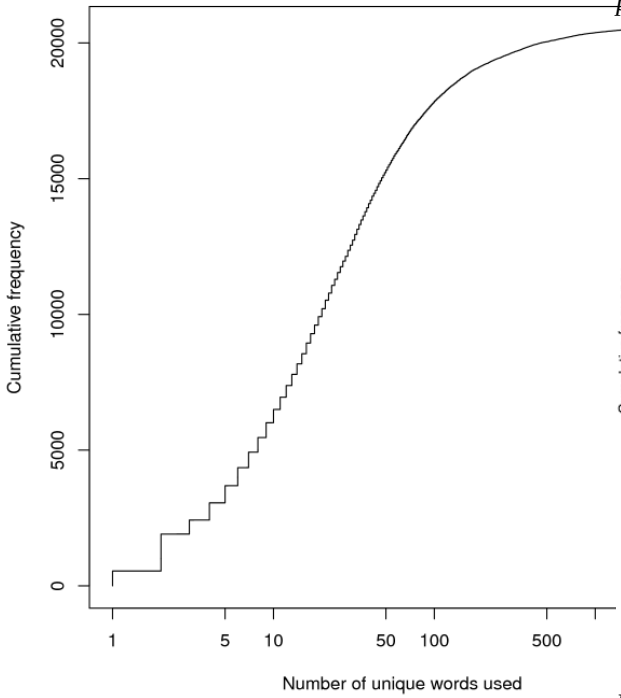


Figure 7: The number of different words used in searches for a given file

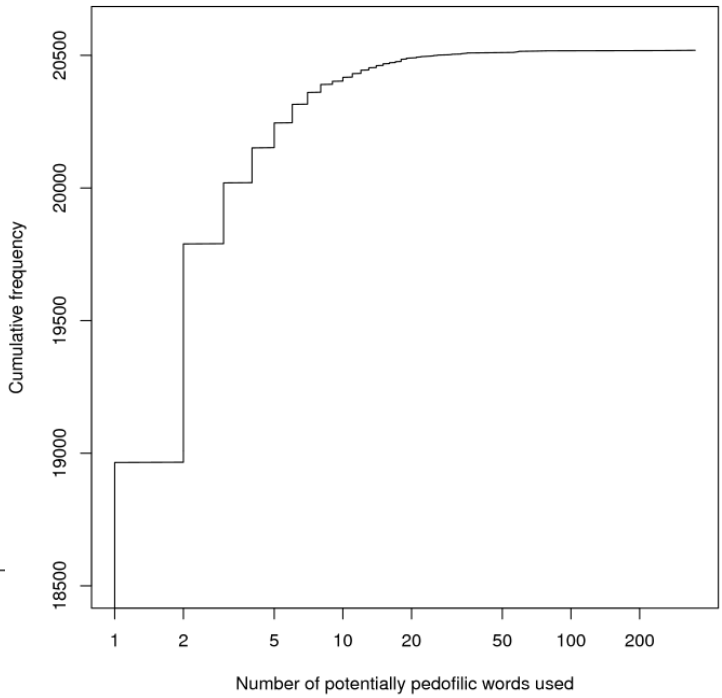


Figure 8: The number of potentially paedophilic words used in searches for files where at lest one potentially paedophilic word used

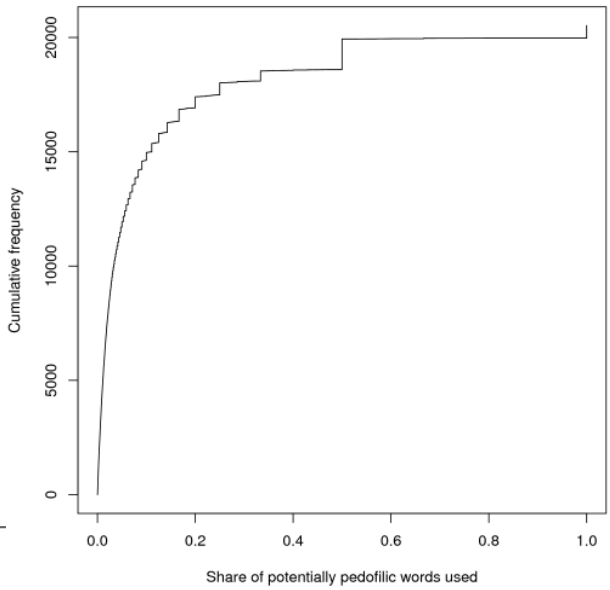


Figure 9: The share of potentially paedophilic words used in searches for files where at lest one potentially paedophilic word used

We also checked what share of all words used to search for a file these potentially paedophilic words represent. The distribution of shares (ratios) is presented in Figure 9 for all words and in Figure 10 for only unique words. We can see that in a about 2400 files these words can represent more than 30% of all words used.

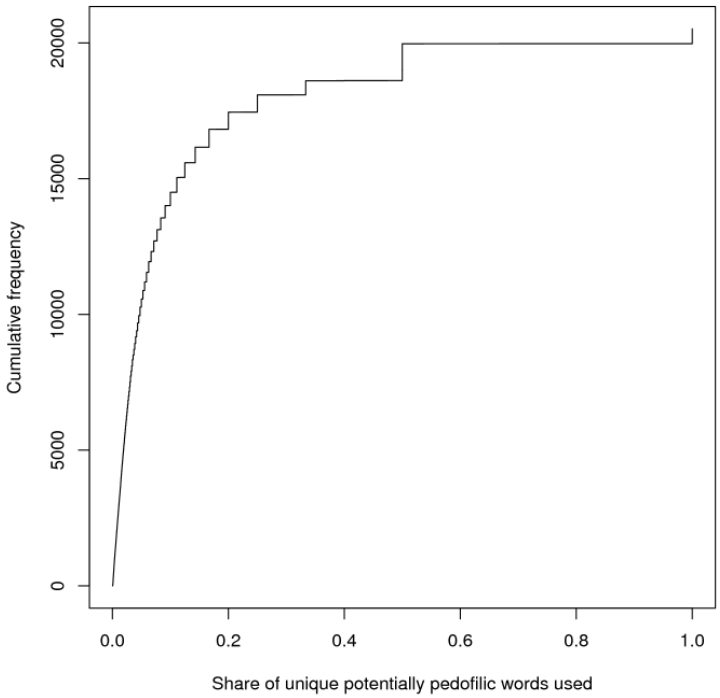


Figure 10: The share of unique potentially paedophilic words used in searches for files where at lest one potentially paedophilic word used

Of the 8991268 files for which the data were available, we had information on which IPs have them for only 5647683 files. Of those 5647683 files, only 334 were offered by IPs that were based on keywords used in their searches classified as "potentially paedophilic". Interestingly, each file was offered by at most one IP that was classified as "potentially paedophilic" and in most cases several other IPs. The following statistics are computed only on those 334 files.

The distribution of the number of IPs that had files were exactly one IP was potentially paedophilic is shown in Figure 11. As we can see, most of those are also hosted by other IPs and therefore should not be termed paedophilic.

The potentially paedophilic IPs have used from 1 to 4 (most of the 4) different potentially paedophilic words to search for files, as is shown in Table 8.

Number of different paedophilic words used	Frequency
1	110
2	93
4	131

Table 8: Average number of different paedophilic words used by IPs that used them and hosted files.

Interestingly the distribution of rations of unique potentially paedophilic words divided by all unique words used contains exactly the same frequencies with an exception that the frequency for 2 in Table 8 is split in two different classes in Table 9, indicating that all these files might be hosted by only 5 different IPs.

Share of unique paedophilic words used	Frequency
0.055556	40
0.057971	61
0.166667	70
0.181818	93

Table 9: Average share of unique paedophilic words used among all unique words used by IPs that used them and hosted files.

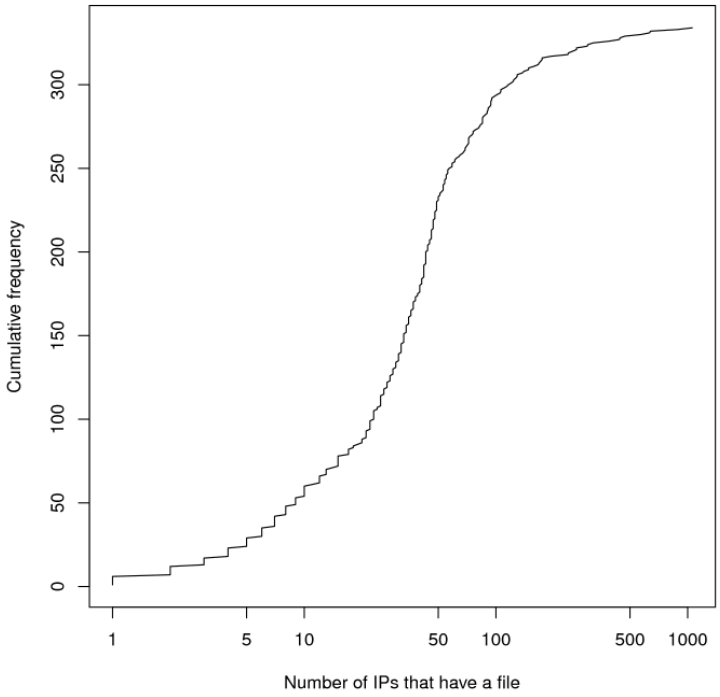


Figure 11: The number of IPs that have hosted a file for files that were hosted by exactly one potentially paedophilic IP

3.5 Identification of contaminated IPs

Similar procedure as for files was also repeated for IPs. In the data analyzed there were 11537290 IPs, of which we had data about keywords they used in searches for 11250750 IPs. Of these 11250750 IPs for which we have data on keywords used, the potentially paedophilic words were used in searches for 20751 of them. The following statistics are computed only on those 20751 IPs.

In Figures 12 and 13 we can see the distribution of number of (unique/different) words used in searches by IPs for which have also used at lest one potentially paedophilic word. The maximum number of different words used by an IP (that also used at least one paedophilic word) is 7810, while the maximum number of non-unique words is 52181.

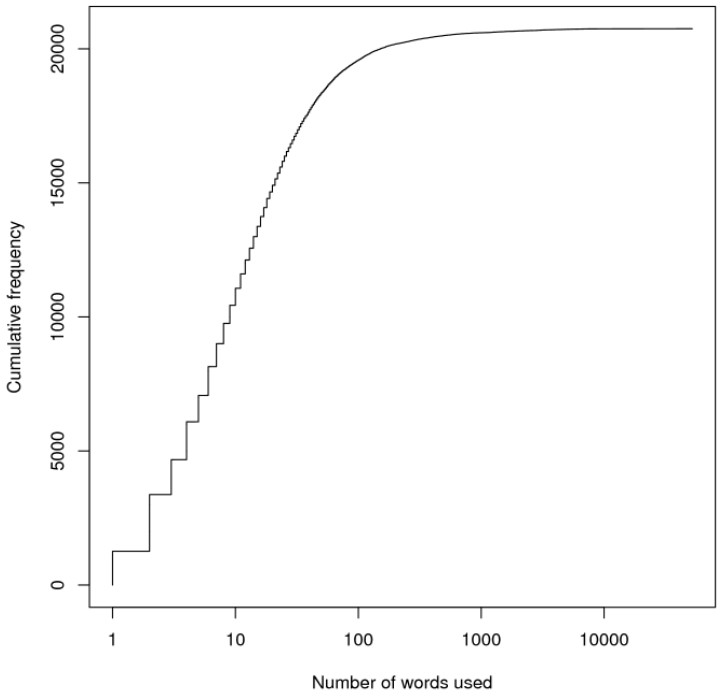


Figure 12: The number of words used in searches by a given IP

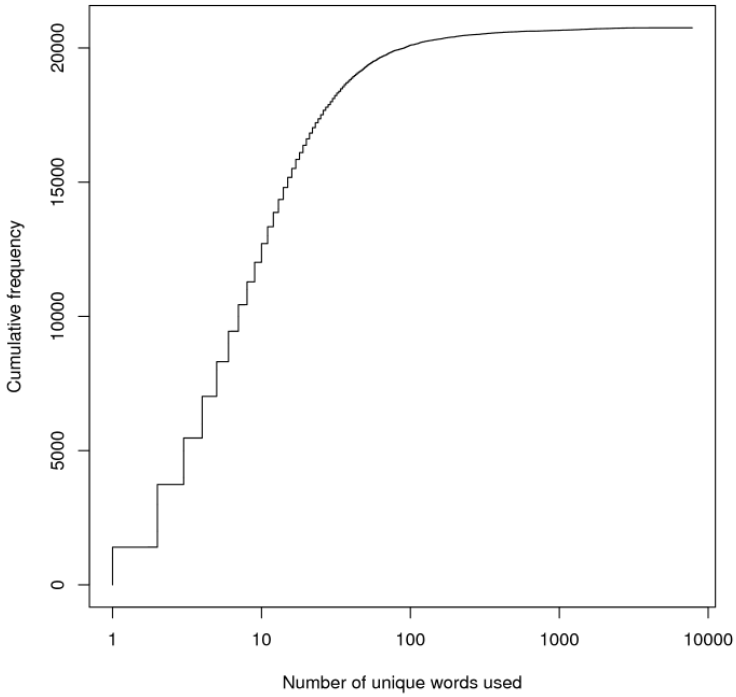


Figure 13: The number of different words used in searches by a given IP

The distribution of the number of potentially paedophilic words and unique paedophilic words is shown in Figures 14 and 15. The maximum number of potentially paedophilic words used by IPs is 285, while the maximum number of unique potentially paedophilic words is 14. However, most of the IPs that did use potentially paedophilic words used only one or two different words.

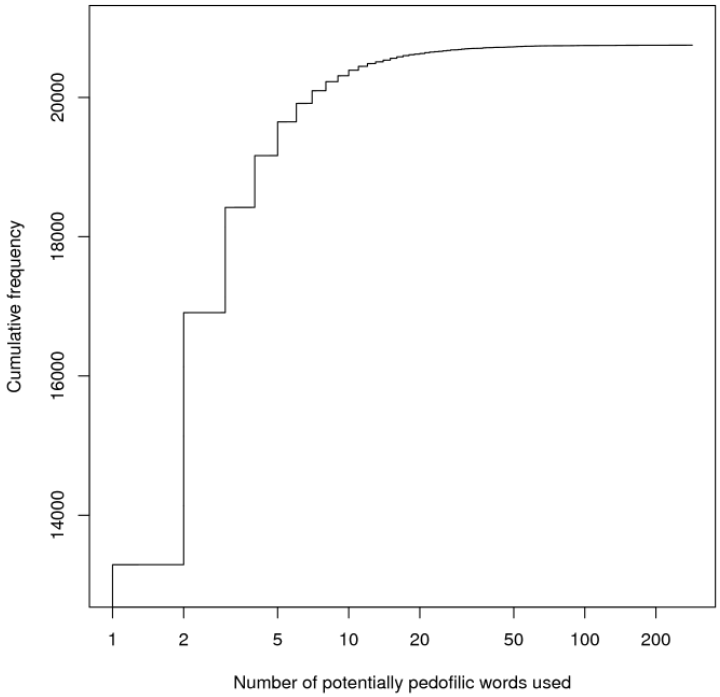


Figure 14: The number of potentially paedophilic words used by IPs in searches

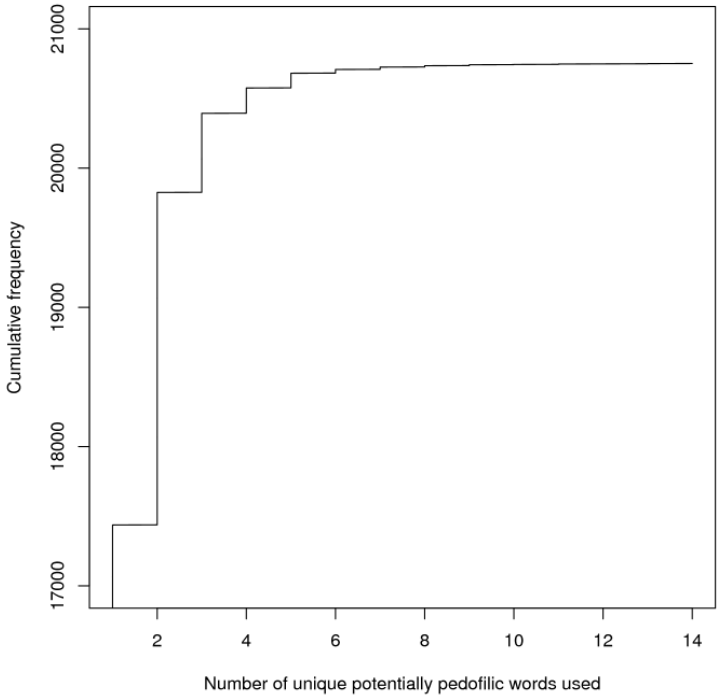


Figure 15: The number of unique potentially paedophilic words used by IPs in searches

Again also checked what kind of share of all words used to search by an IP these potentially paedophilic words represent. The distribution of shares (ratios) is presented in Figure 16 for all words and in Figure 17 for only unique words. We can see that in a few 1000 IPs these

words can represent significant portions. However, we should take into account most of these large shares (ratios) occurred for IPs that used only a few keywords (made only a few searches). This also evident from the large “jumps” of the curve at 1, 1/2, 1/3, 1/4, ...

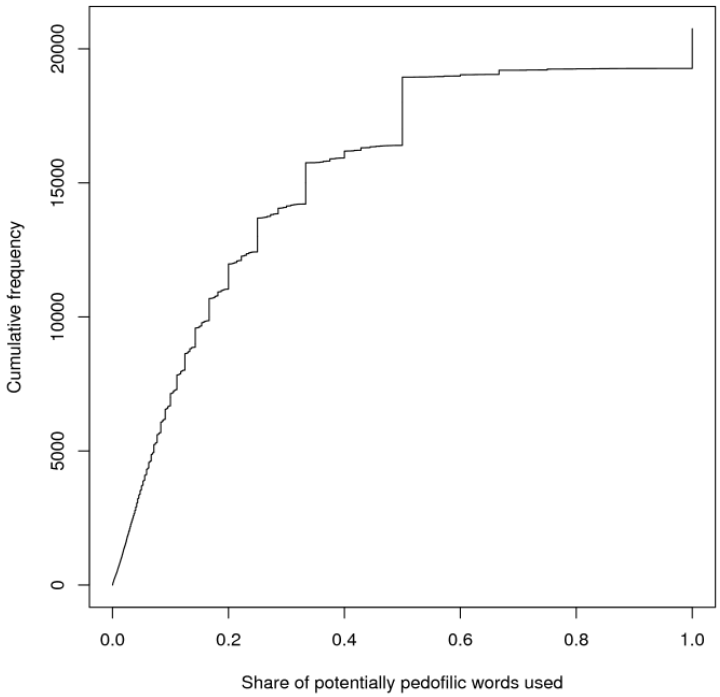


Figure 16: The share of potentially paedophilic words used in searches by a given IP

Of the 12270786 IPs for which appeared in the data, we had information on files that IPs have for only 1022786 IPs. Of those, 368543 IPs had in possession at lest one potentially paedophilic file. The following statistics hold for those 368543 IPs. In Figure 18 the distribution of the number of files that an IP has is shown and in Figure 19 the number of potentially paedophilic files. Some IPs had up to 163 potentially paedophilic files, while more than 18000 of them have 5 or more potentially paedophilic files. The ratio (share) of potentially paedophilic files is shown in Figure 20. While most IPs have a small share of potentially paedophilic files, more than 3000 have more than 50% of potentially paedophilic files.

To check how likely it is that potentially paedophilic files are really paedophilic, we aslo present the average number of unique potentially paedophilic words in potentially paedophilic files that an IP has on Figure 21 and the average share of unique potentially paedophilic words among all unique words in Figure 22. We can see that most IPs have files with on average only a one potentially paedophilic word. Only a few IPs have potentially paedophilic files with a significant average share of potentially paedophilic. E.g., only 11 IPs have files that have more than 50% unique paedophilic words.

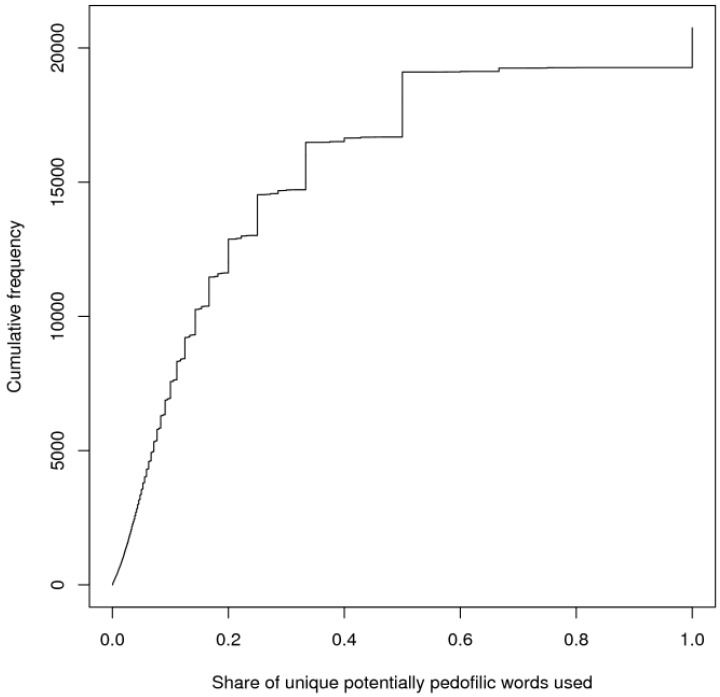


Figure 17: The share of unique potentially paedophilic words used in searches by a given IP

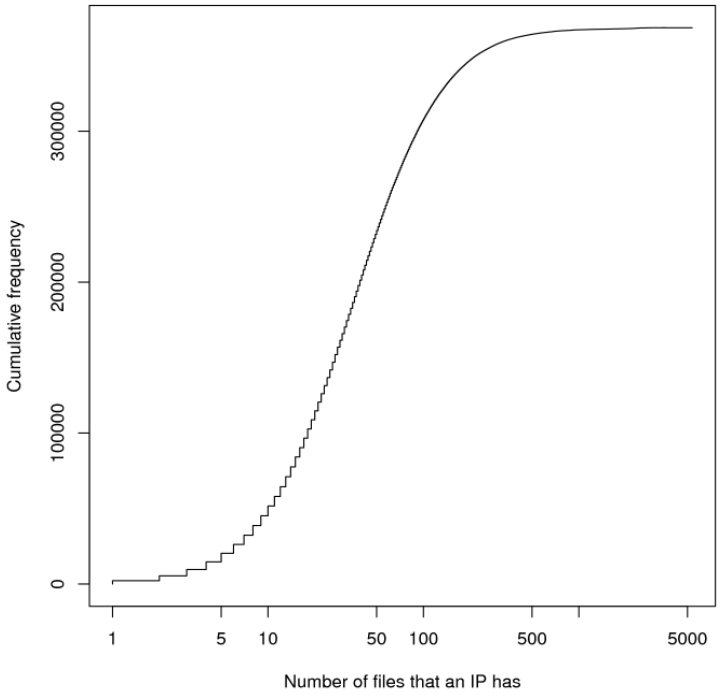


Figure 18: The number of files that IPs hosted

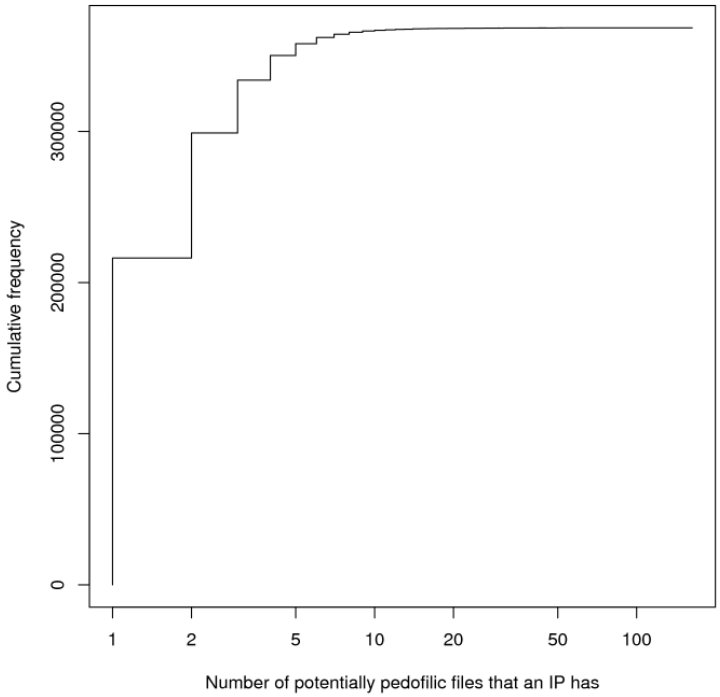


Figure 19: The number of potentially paedophilic files that IPs hosted

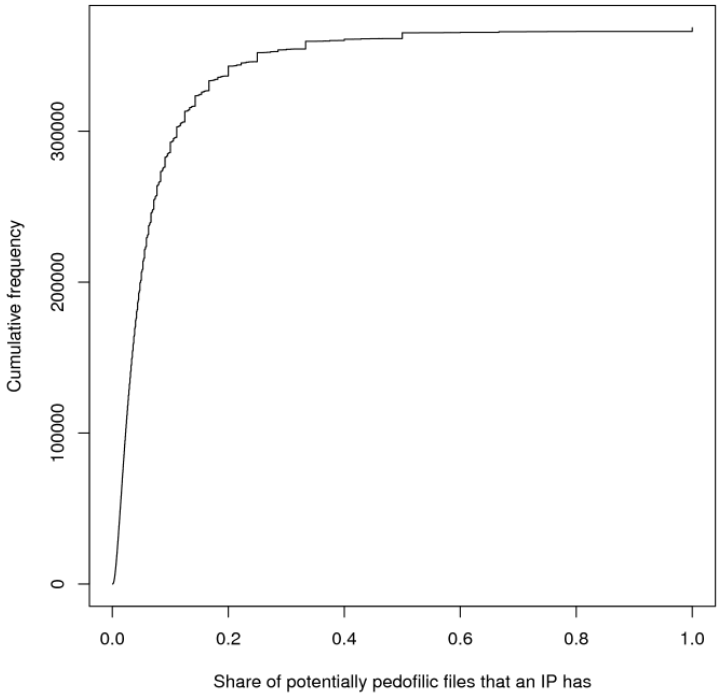


Figure 20: The share of potentially paedophilic files that IPs hosted

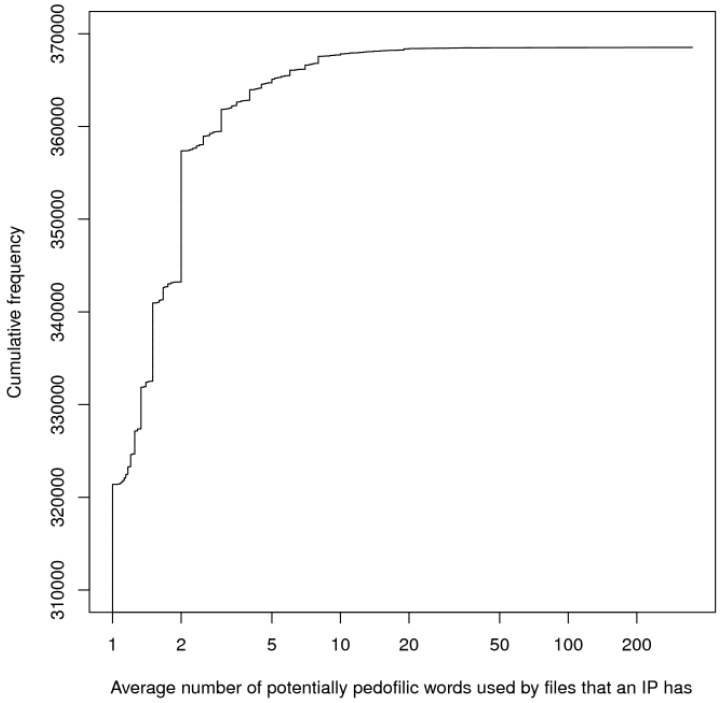


Figure 21: Average number of different paedophilic words used by IPs that used them and hosted files.

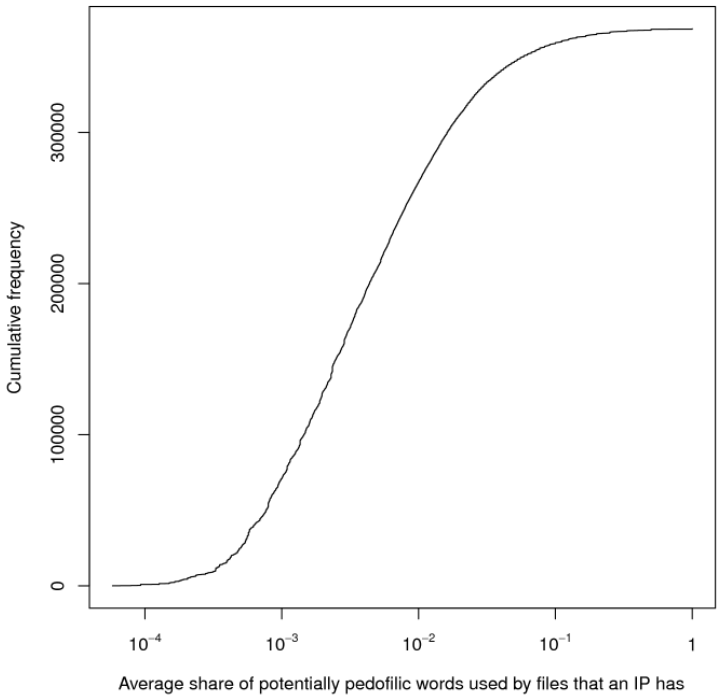


Figure 22: Average share of different paedophilic words used by IPs that used them and hosted files.

3.6 Key demographics summarised

Here we are summarizing the key demographic of our datasets:

Files:

- Number of all files: 8 991 268
- Number of files with data on words used to find them: 6 819 038
- Number of files that were found using paedophilic keywords: 20 519 (for 20 228 only one unique word was used to find them, for 272 files 2 and only 19 files 3 different words were used)
- Number of files with at lest one known host (IP): 5 647 683
- Number of files hosted by "paedophilic" IPs: 334
- Number of all IPs: 11537290

IPs:

- Number of all IPs: 12 270 786
- Number of IPs with data on words used in searches: 11 250 750
- Number of IPs that used paedophilic words in searches: 20 751
- Number of hosts (IPs that host files): 1 022 786
- Number of hosts (IPs) with at lest one potentially paedophilic file: 368 543

The ratios IPs to files or vice versa:

- All IPs to all files: 12 270 786 / 8 991 268 = 1.4
- All hosts to files with at lest one known host: 5 647 683 / 1 022 786 = 5.5
- Hosts with at lest one peodphilic file to paedophilic files: 368 543 / 20 519 = 18.0
- Hosts that used paedophilic words to files hosted by them: 20 751 / 334 = 62.1

4. Strings and keywords related to contaminated files/search IPs

After having identified potentially paedophilic IPs and files, we checked which keywords were used in related queries. Here we took into account only IPs/files that were initially identified as "potentially" paedophilic based on the keywords used in corresponding queries.

4.1. Keywords related to contaminated search IPs

In Table 10, the keywords appearing most frequently in contaminated queries – the queries where at least one previously known keyword was used) made by potentially paedophilic IPs (those that used at least one paedophilic word in some of their queries) are presented. We can see that, of course, known contaminated words rank high here.

Word	Freq
Pthc	**19076**
The	12303
Pedo	**9169**
Mpg	8362
Avi	8017
Jpg	6947
ita	6596
of	5505
2	5005
in	4672
sex	**4605**
a	4340
de	4303
and	4030
la	3906
2006	3801
boy	**3777**
girl	**3642**
1	3407
i	3311
s	3292
e	3064
anal	**3062**
mp3	2965
live	2890

Table 10: The 25 most frequent word in queries made by contaminated IPs that used at lest one "potentially paedophilic" word sorted by frequency of use

In Table 11, the frequencies of appearances of contaminated words in these queries made by contaminated (potentially paedophilic) IPs are presented.

Word	Freq
pthc	19076
pedo	9169
ygold	2605
hussyfan	2259
pedofilia	2117
lolitas	2058
ptsc	1768
lolitaguy	759
childlover	649
babyshivid	555
qwerty	340
kinderficker	320
nymphets	307
lolitasex	298
KIdzilla	269
kdquality	241
zoophilia	218
childporn	215
pedoland	209
kiddie	204
paedophile	167
ddoggprn	126
pedofilo	122
pedophilia	119
childfugga	95
pedos	91

Table 11: The number of times each "potentially paedophilic" word was used in queries made by IPs that used at lest one "potentially paedophilic" word sorted by frequency of use

As the absolute values are not a very good indication of the tendency of the word to be used in searches for paedophilic content, we for also computed for each word the share of its appearances in searches by potentially paedophilic IPs / leading to paedophilic files to better estimate its likelihood of being used for paedophilic purposes.

4.2. Keywords related to contaminated files

In Tables 12 and 13 similar statistics are presented in all queries related to contaminated files, i.e. those that were found in queries containing potentially paedophilic words.

word	**freq**
the	22900
de	18001
la	12073
2	10373
of	10317
fr	10169
lolitas	**10073**
a	8937
ita	8433
2006	7082
mst	6966
3	6640
el	6423
pthc	**5898**
i	5777
le	5659
e	5319
dvd	5269
me	5110
s	4988
in	4895
pc	4752
prison	4690
break	4609
1	4488

Table 12: The 25 word that appear most frequently in queries that included files that were also found by at lest one "potentially paedophilic" word sorted by frequency

word	
lolitas	10073
pthc	5898
pedo	2850
lolitaguy	1252
ptsc	492
hussyfan	392
qwerty	165
nymphets	131
zoophilia	88
lolitasex	86
babyshivid	53
pedofilia	47
ygold	46
ddoggprn	37
kdquality	32
kidzilla	17
childlover	12
kiddie	11
pedophilia	4
pedoland	4
kinderficker	1

Table 13: The number of times each "potentially paedophilic" word was used in queries that included files that were also found by at lest one "potentially paedophilic" word sorted by frequency

5. Networks of IPs, files and keywords

The basis for the analysis presented below is a 3 mode network with the following ties:

- IPs – Files : which files were hosted by each IP
- IPs – Words : which „paedophilic“ words were used (in queries) by IPs
- Files – Words : which „paedophilic“ words were used in queries that lead to files

The network is therefore composed of:

- 389291 IPs – These are the IPs that either searched for “paedophilic” words or had files that were found by queries containing “paedophilic” words. Of theose, 20751 have used “paedophilic words in queries”
- 20847 Files – These are files that were either found using queries containing “paedophilic” words or were hosted by IPs that used “paedophilic” words in queries. Most of those (20519) were found using queries containing paedophilic words.
- 26 Words that were selected for their paedophilic use.

The whole network will be analyzed by type of ties (the three noted above).

As even this largest component is too large to draw, we have further reduced it by keeping only those files that were connected to at lest two words. Such a network has than only 317 units (or 303, if we do not count the words that are not connected to any files in this network – we kept them in as a reminder that they were used). It is presented in Figure 23. The use of this network is however limited. Practically the same information is presented in a clearer way in the form of one-mode network of words based on co-occurrences in files.

Based on this network two one-mode networks can be created:

1. A network of words that were used to find the same file
2. A network of files that were found using the same words

The two-mode network of files and Words contains 20519 files (as the rest are not connected to words) and 21 words (5 words were not used in any query that led to a file). The whole network contains 20829 ties (edges) and is as such too large to draw. The network is composed of only one (“giant component”) is contains most of the units (20314), while the rest contain only one word and the files that are connected to only that word.

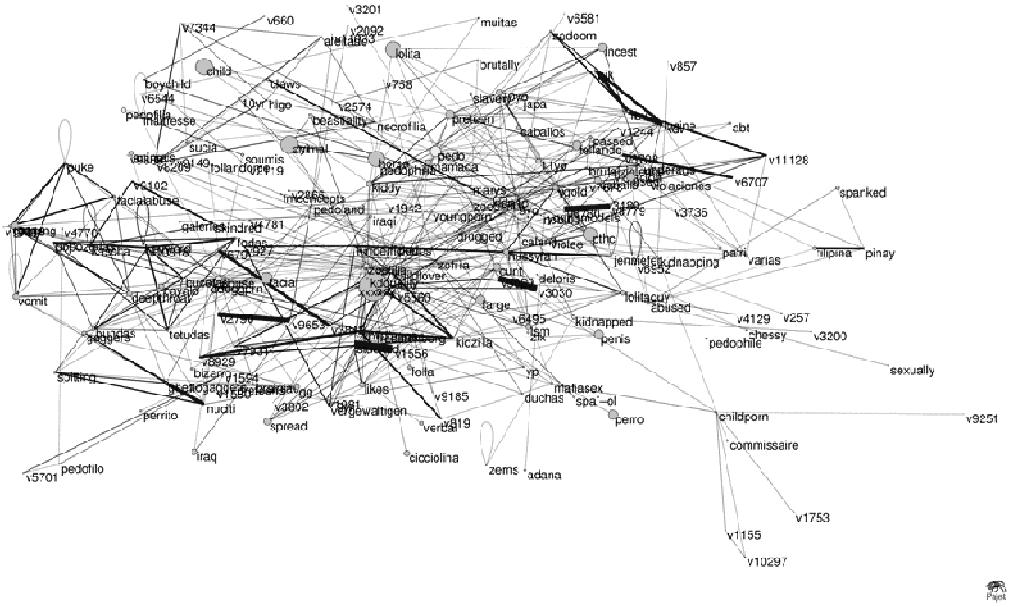


Figure 23: Two-mode network of files (yellow vertices - (only files with ties to at lest 2 "paedophilic" words are shown) and "paedophilic" words

5.1 Analysis of the networks

We perform here actions (a) and (b) from the SECOND step outlined in the Introduction.

5.1.1 Additional keywords from contaminated search IPs

The shares of appearances of words in searches by potentially paedophilic search IPs are presented in Table 14 for words appearing more than 10 times and are in more than 50% used by "potentially paedophilic" IPs.

rank	word	share	freq	rank	word	share	freq
1	madebyarkh	0.82	11	35	cbaby	0.58	202
2	nakie	0.81	170	36	8y	0.57	94
3	ssap	0.81	31	37	extremep2p	0.57	28
4	invideo	0.80	107	38	rbv	0.57	260
5	totp2	0.79	85	39	vtcap	0.56	163
6	ekus	0.77	13	40	totalupdate	0.56	16
7	reelkiddymov	0.76	119	41	zadoom	0.56	322
8	21min	0.75	12	42	9yo	0.56	1355
9	5y	0.74	61	43	7yo	0.56	879
10	halyavapictures	0.72	72	44	9y	0.56	113
11	ch1	0.72	18	45	8yo	0.55	966
12	beerbarrel	0.72	32	46	qqaazz	0.55	485
13	jeffz	0.71	35	47	ezik007	0.55	29
14	stucked	0.69	16	48	xvcd	0.55	29
15	abner	0.68	142	49	602	0.55	71
16	babyj	0.67	601	50	nudisten	0.55	42
17	cduk	0.66	114	51	sedna	0.55	64
18	datacd	0.65	20	52	10min	0.55	11
19	alysia	0.65	79	53	pae	0.55	200
20	nobull	0.63	82	54	motivational	0.54	57
21	fallenangelfuns	0.62	32	55	bandler	0.54	456
22	nuciti	0.62	45	56	newestmp3s	0.54	13
23	lolalover	0.62	50	57	videodead	0.54	41
24	senatorinfo	0.62	21	58	rizmastar	0.53	68
25	liluplanet	0.62	515	59	kingpass	0.53	915
26	lordofthering	0.61	239	60	aist	0.53	93
27	thoroughly	0.61	77	61	eurololita	0.53	114
28	samal	0.60	397	62	ura101	0.52	168
29	kinofack	0.60	210	63	euman	0.52	125
30	phx	0.59	27	64	mse	0.52	29
31	Jenniefer	0.59	97	65	5yo	0.51	751
32	Wixar	0.59	63	66	infolowy	0.51	43
33	Uvs	0.58	161	67	nudism	0.51	518
34	Soperedi	0.58	74	68	nablot	0.51	391

Table 14: The words that appear in the analyzed data more than 10 times and are in more than 50% used by "potentially paedophilic" IPs sorted by the share of their appearances in searches made by "potentially paedophilic" IPs

Considerable part of these new keywords could be immediately confirmed with a simple web search (e.g. madebyarkh).

Further checking for some of those words in Urban dictionary (http://www.urbandictionary.com/) can show for example for “kingpass” that it is also used for tagging paedophilic content. Therefore, all these words should be substantially checked.

Interestingly, the distributions for potentially paedophilic files and search IPs are dramatically different. While the most frequent words in queries made by potentially paedophilic IPs are consistent with searching for paedophilic content and the most frequent word is definitely a paedophilic word, this can not be said for paedophilic files, or files named (most likely wrongly) potentially paedophilic.

5.1.2 Additional keywords from contaminated files

The shares of appearances of words in searches that found potentially paedophilic files are presented in Table 15 for words appearing more than 50 times and are in more than 70% used to find "potentially paedophilic" files (the criteria for inclusion are here stronger to limit the number of words to a manageable number). As before, all these words should also be checked, although based on results presented in Tables 12 to 13 we assume that there will be less paedophilic words here.

If we set the benchmark to 50% we obtained 58 new key words; many of them being directly recognized as paedophilic by simple web search (e.g. reelkiddymov).

Surprisingly, there is little overlap with the keywords in the table 14.

rank	word	share	freq	rank	word	share	freq
1	schoo	1.00	78	30	shadoks	0.80	1061
2	t4c	1.00	55	31	vizi	0.80	3495
3	reelkiddymov	1.00	119	32	coreavc	0.80	80
4	cennet	1.00	77	33	sunt	0.79	295
5	straat	1.00	61	34	sedna	0.79	64
6	velos	1.00	53	35	galã	0.79	461
7	amentes	0.96	164	36	7y	0.79	57
8	mst	0.95	559	37	gagoule	0.78	968
9	komórka	0.95	2020	38	2205	0.78	464
10	contagium	0.93	117	39	yasuda	0.78	55
11	u50	0.90	98	40	gomon	0.77	162
12	agepito	0.89	89	41	antonieta	0.77	270
13	brokes	0.88	282	42	jop	0.76	190
14	komorka	0.88	686	43	pozostal	0.76	504
15	splendore	0.87	51	44	turra	0.76	69
16	shinedoe	0.87	98	45	61101	0.75	488
17	mrasche	0.86	333	46	lewa	0.74	241
18	cashback	0.86	79	47	derapage	0.74	231
19	tomton	0.85	82	48	16386	0.74	555
20	leod	0.84	164	49	czlowiekiem	0.74	646
21	akoustic	0.84	68	50	middleman	0.72	540
22	primi	0.84	10007	51	5y	0.71	61
23	aquisizione	0.83	71	52	rous	0.71	73
24	shinsengumi	0.83	101	53	tihij	0.71	86
25	ved	0.82	92	54	vieu	0.70	174
26	3615	0.82	133	55	phee	0.70	2022
27	shortbus	0.82	2746	56	papiez	0.70	710
28	jaggets	0.81	293	57	hokkabaz	0.70	1113
29	alaskan	0.80	122	58	baci	0.70	3773

Table 15: The words that appear in the analyzed data more than 50 times and are in more than 70% used to find "potentially paedophilic" files sorted by the share of their appearances in searches that found "potentially pedohilic" files

5.1.3 A network of keywords and files

The network of keywords used to find the same file is small as we selected only 26 “paedophilic” words and relative sparse network, as most files were found using only one word. The tie values as cosine similarities among words computed based on the two-mode network presented in the above. This network is presented in Figure 24. This network actually presents a similarity matrix. As such it can be simply converted to dissimilarity matrix and used as an input to clustering. The result of hierarchical clustering using Ward's method is presented in Figure 25.

Especially the graph in Figure 25 shows us that most words are not connected (there is no files that would be searched for using both words). There are however a few pairs of words that are relatively often used together, the most notable one being “querty” and “ddoggprn”.

The network files related to the same keywords does not seam useful. The whole network 20751 units (files) and 140 mio ties. As such, it is too large to be analysed. The tie values represent the number of same words used to find a two files. Most of the ties have a value of 1, while 11882 have value 2 and 133 a value 3. As such network contains practically the same information as the two-mode network presented above, we are not analysing it further.

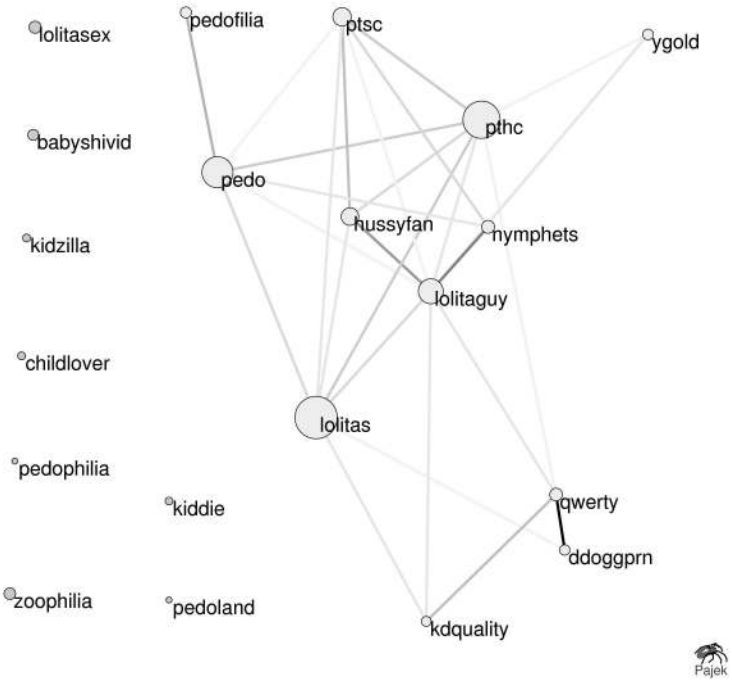


Figure 24: One-mode network of “paedophilic” words based in their ties to files. The tie width is proportional to the square root of the cosine similarities and the size of the vertices to the square root of the frequencies.

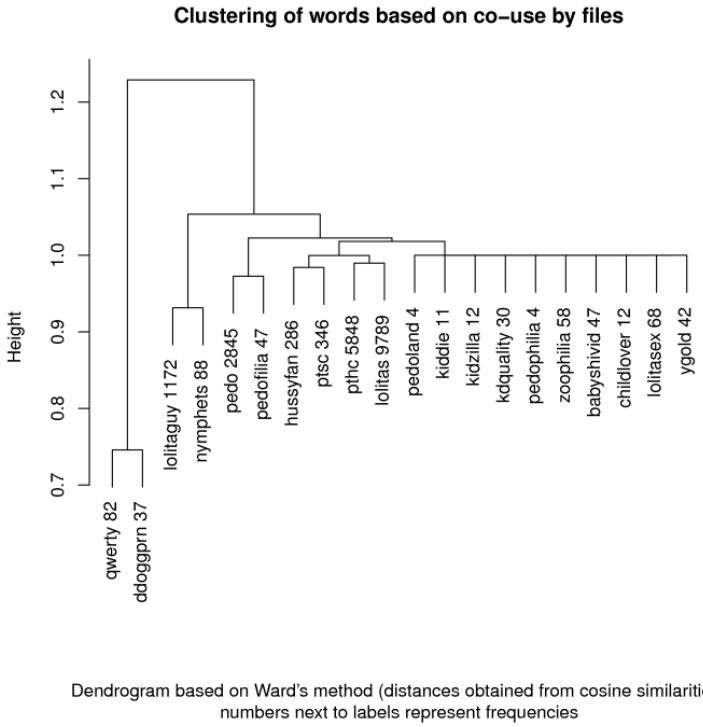


Figure 25: Hierarchical clustering of "paedophilic" words based on cosine similarities computed based on the files-words two-mode network

5.2 Analysis of ties among IPs and keywords

5.2.1 Two-mode network of IPs and keywords

The two-mode network of IPs and words contains essentially 20751 IPs (as the rest are not connected to words) and 26 words and is as such too large to draw. The whole network contains 25704 ties (edges). The network contains only one component (there are no disconnected units/groups). As this network is to large to draw we present in Figure 26 a network without the IPs that have used only one "paedophilic" word in their searches. As before, the relevant information is probably better presented through a one-mode network of words (where the ties represent co-use of words by IPs).

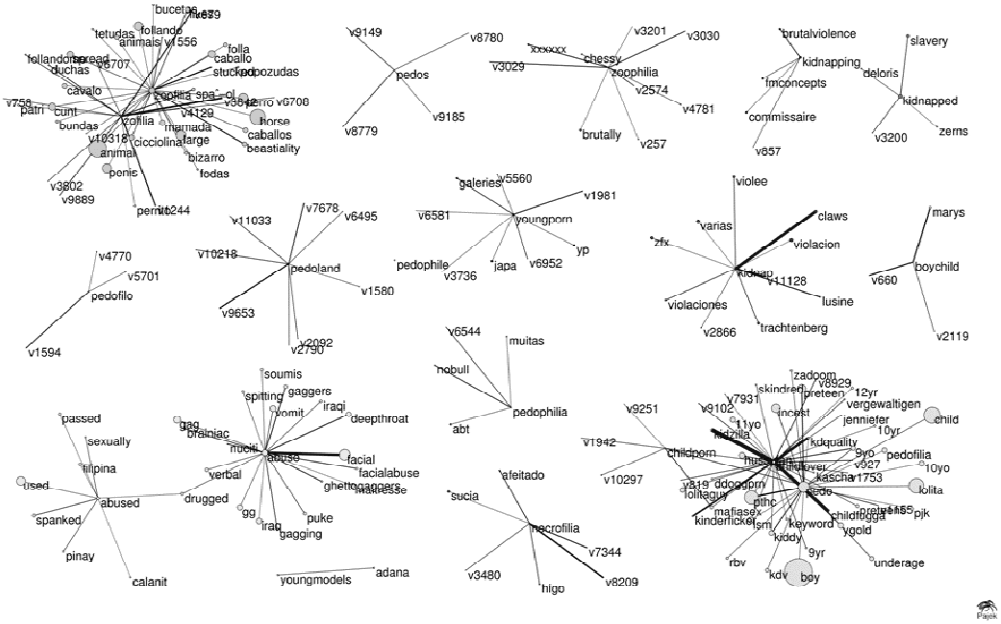


Figure 26: Two-mode network of IPs and "paedophilic" words. The size of the vertices representing IPs is 0, so we can only observe ties connected to them. Only IPs with ties to at least two "paedophilic" words are drawn.

Based on this network two one-mode networks can be created:

1. A network of words that were used by the same IPs
2. A network of IPs that used the same words

5.2.2 A network of words that were used by the same IPs

This is a small network as we selected only 26 words, as IPs use more different words in searches, this network is much denser. The ties values represent cosine similarities based on files. The network is presented in Figure 27. What we can see here is that the term “pthc” (pre-teen hardcore” seams to be central (similar to most words and connecting them), indicating that the words were selected correctly. There are a few words that do not seem to fit (e.g. pedophilia/paedophile, pedos, zoophilia, lolitasex, ...).

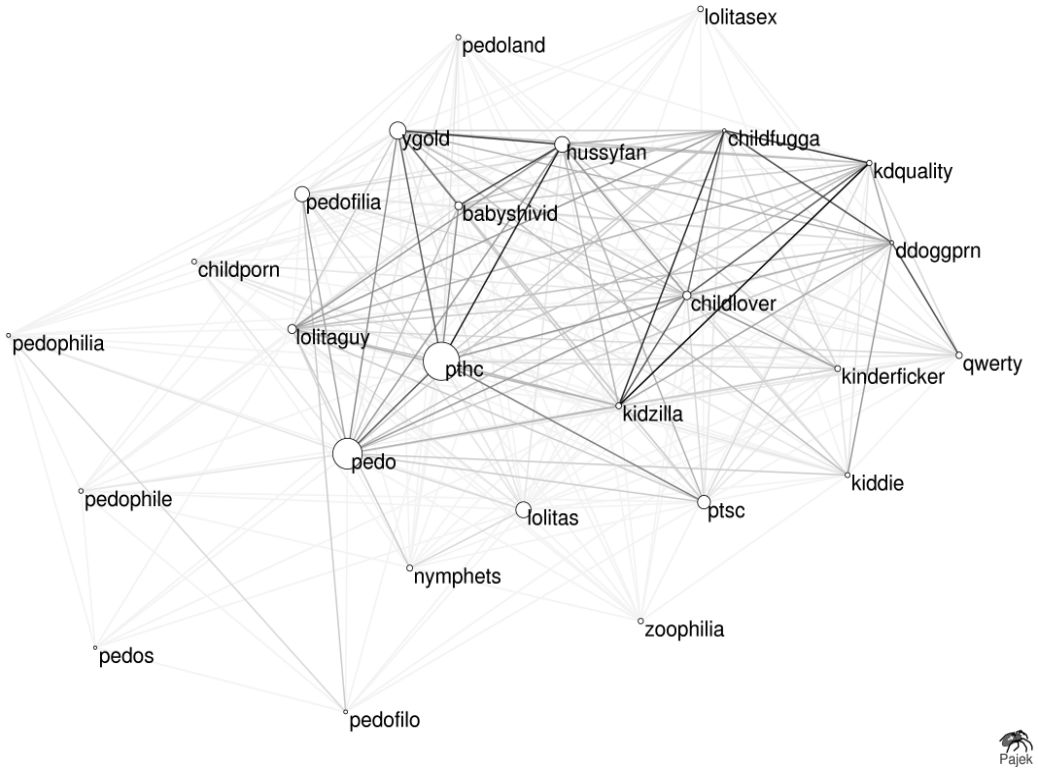


Figure 27: One-mode network of “paedophilic” words based in their ties to IPs. The tie width is proportional to the the cosine similarities and the size of the vertices to the frequencies

These similarities were also used as an input to hierarchical clustering using ward's method. The results are presented in Figure 28.

5.3 Analysis of ties among IPs and files

5.3.1 Two-mode network of IPs and files

The two-mode network of IPs and files contains 368,545 IPs (as the rest are not connected to files –t hey do no host them) and 16742 files (the rest have no ties to IPs) and is as such too large to draw. The whole network contains 684042 ties (edges). The network contains 764 components, however most of the units (382287 files and IPs) are in the giant component (there are also two components of size 71 and 62, while the rest contain 22 units or less. As we are especially interested in files/IPs associated with "paedophilic" words, we are also using in this analysis the information on the number of "paedophilic" words associated with a given file/IP or the share of "paedophilic" words among all words associated with a given file/IP.

Our analysis showed that of the 20751 IPs that used "paedophilic" words in their searches, only 5 hosted file(s) (that is 0,02%). Of those 5, only 3 hosted "paedophilic" files. In comparison, out of 11 mio IPs that made at lest one search, only 2750 IPs have hosted at lest one file (that is again 0,02%). Therefore, the IPs that search for files usually does not host them (and vice versa). This makes automatically means that in our IPs – files network, we will have very little "paedophilic" IPs, as to be names "paedophilic", they have to *search* using "paedophilic" words, while most hosts do not search at all.

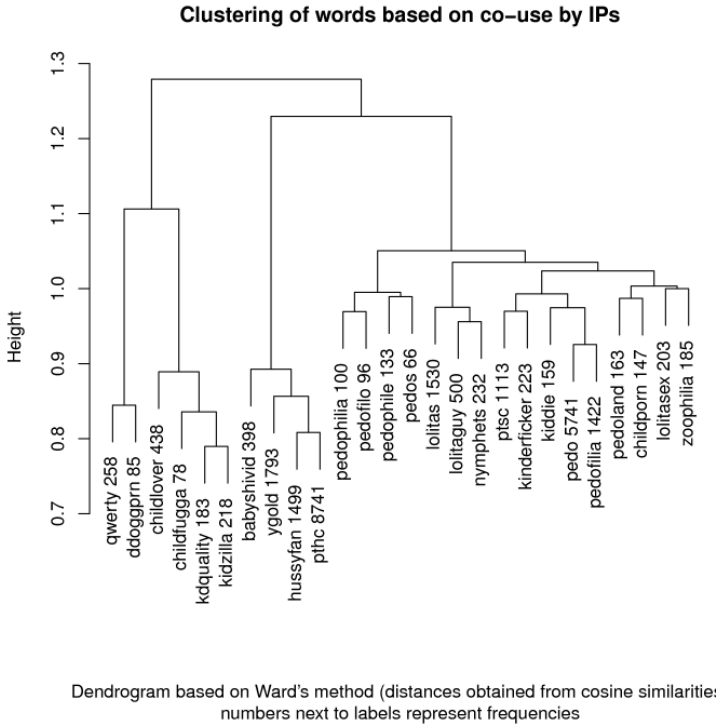


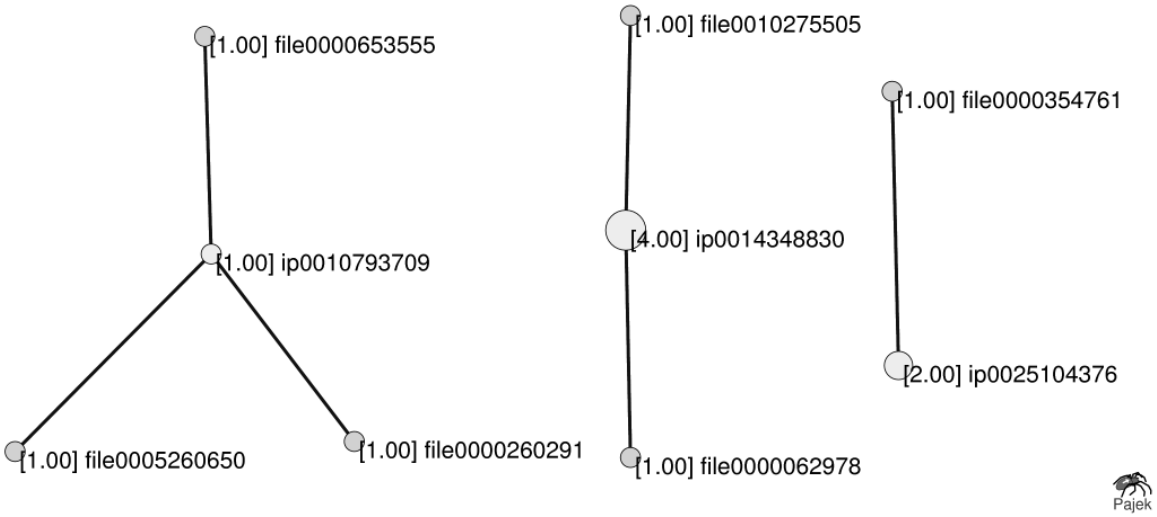
Figure 28: Hierarchical clustering of "paedophilic" words based on cosine similarities computed based on the IPs-words two-mode network

We first extracted only the files/IPs (directly) associated with "paedophilic" words. After we excluded files/IPs with no ties, only 3 IPs and 6 files remained. The resulting network is shown in Figure 29. This figure was further extended in Figure 30 by adding the "paedophilic"words associated with these files/IPs.

While the giant component of the whole two-mode (files – IPs) network is too large to be shown, we can easily draw some smaller components. Most of the components are

star-like, that is either composed on one IP and several files that this IP has or of one file and several IPs that host this file. Relatively numerous structures are also a few IPs that host the same files (or vice versa, which is actually the same) or something similar (only a part of the files are hosted by all IPs). In Figure 31 we are presenting some of the more interesting shapes. The yellow vertices are IPs and the green ones are files. The size of the vertex is determined by the share of "paedophilic" words used (in searches by IPs or in searches that led to a file for files). I we see just a line without a vertex on the end, that means that that vertex (file or IP) did not use any "paedophilic" words. As files are connected only to IPs and vice versa, we know the type of the vertex to be the opposite to the vertices to which it is connected. At lest one vertex for each tie has some "paedophilic" words associated with it, as this was the condition for inclusion in the network. We can see that in the components show in all but one (there were two in all "small" components") component the "contaminated" unit is a file.

As mentioned earlier, the giant component is too large for practical analysis. For each IP/file we computed the share of "paedophilic" words among all words associated with it and also the average of such statistic of its neighbours (either files or IPs). Then we computed the mean of these two values. We then excluded the units (files and IPs) which had the mean below 0.1 to get a much smaller and manageable network while hopefully keeping the most "paedophilic" parts of the network. This reduced network was composed of 1499 components, 50 of which had 30 or more units. This also includes one giant component with 53051 units. We again first explored the remaining 49 components with from 30 to 77 units. These are presented in Appendix 1, where the size of the vertex is proportional to the share of "paedophilic" words among all words. We can notice an interesting pattern – all "paedophilic" units are files (green). This is expected due to small number of



"paedophilic" IPs that host files.

Figure 29: Files and IPs associated with "paedophilic" words

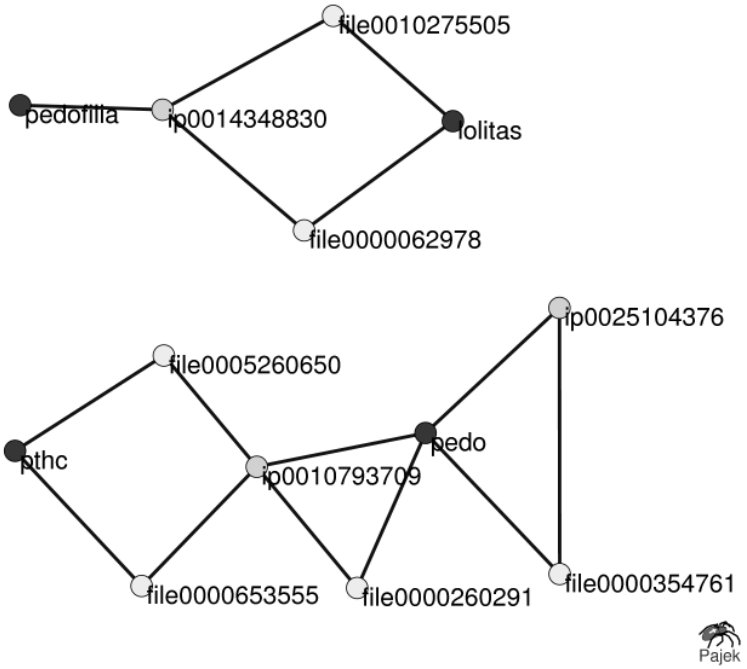


Figure 30: Files and IPs associated with "paedophilic" words and the associated words

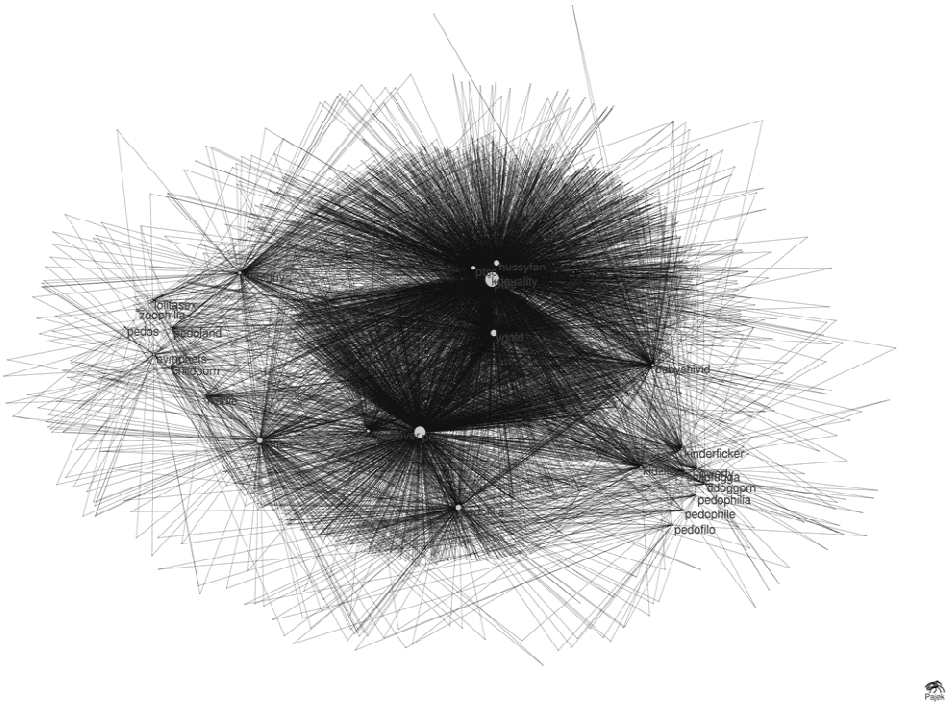


Figure 31: A selection of smaller components in the file (green) – IPs (yellow) two-mode network. The size of the network represents the share of the "paedophilic" words among all words associated with a given file/IP.

As the giant component, even on this reduced network is way too large to be effectively analysed. The islands (presented in Appendix 2) are not much different from the components above (which is not surprising, as vertex islands algorithm can be tough of as a procedure for "smart" selection of threshold values for vertex cuts an then extracting components. There are however some especially interesting islands, in particular an island with both "paedophilic" IP and files (although not directly connected). For this and some other islands we produced separate plots that will also include "paedophilic" words. These are presented in Appendix 3.

We can also observe numerous configurations where several "paedophilic" files are all hosted by the same IPs. This might indicate that these files and the IPs that host them are somehow connected, if not by anything else by the common "interest" or "content", possibly paedophile. As can be seen from Appendix 3, these files are not necessarily "incriminated" by the same words, indicating that they might have something else in common. However we should check if this is not perhaps some other topic.

5.4 Discussion of the results

The biggest surprise is that very few files and IPs were found as potentially paedophilic based on both criteria (keyword and connections to other files/IPs). This is however not so much surprising if we take into account the specific fact that most of the IPs that search for files do no host them (and vice versa). This is especially surprising, as the eDonkey system is based on the fact that people also share the files that they are downloading.

Another surprise is that for most files that were classified as potentially paedophilic, only one potentially paedophilic word was used to find them. This might indicate that paedophilic content is practically not present, or, that paedophilic files are tagged usually with only one "incriminating" tag to avoid detection by outsiders. However, these files can usually be also found using a relatively large number of other (non-paedophilic, at lest to our knowledge) words. Additional words in the title might also be used to avoid detection.

In total, 20,751 files were found hit by potentially paedophilic words. However most of them were found with only one potentially paedophilic word (some aslo with 2 and 3). Very few files (only 334) are hosted by IPs that were termed "potentially paedophilic" based on the keywords they use in searches and each by only one IP.

In total 20 519 search IPs used paedophilic words, however most of them again used only 1 unique keyword (only 176 used 5 or more, at most 14). A large number of IPs also hosted potentially paedophilic files (368 543), although, as mentioned before, these are not the same IPs as those that search for paedophilic content. There were only 3 "paedophilic" IPs that hosted "paedophilic" files, and all of them together hosted only 5 "paedophilic" files.

However, we have observed numerous network configurations where several "paedophilic" files are all hosted by the same IPs. This might indicate that these files and the IPs that host them are somehow connected, if not by anything else by the common "interest" or "content", possibly paedophilic. Interestingly, these files are not necessarily "contaminated" by the same words, indicating that they might have something else in common.

Additional interesting finding is that the identification of "potentially paedophilic" IPs was more successful than that of "potentially paedophilic" files. We can conclude that contaminated keywords were much more consistent with paedophilic nature for IPs than for words. This is however understandable, as files could be found also by searches containing paedophilic words and non-paedophilic words (and if the non-paedophilic words are responsible for the file being found, this is not really an indication of paedophilic content). On the other hand, someone who uses paedophilic words in his/her searches is most likely really interested in finding paedophilic content.

6. Future work

Possible further research could go into the direction to complete the analysis outlined in the introduction, i.e. the iterations of the full performance of the FIRST and also SECOND step until convergence is reached. This process need to be additionally parameterised, so that stable set of contaminated elements (files, keywords, IPs) will be identified.

Firstly, it would be thus beneficiary to repeat similar analysis with an improved list of potentially paedophilic words, giving them a "score" indicating about the strength of the paedophilic nature. This would allow getting better and more precise evaluation for paedophilic nature also for files and IPs, which may also receive similar scores. The entire process should then iterate until the scores converge as outlines in introduction. Further analysis could be also extended by taking into account additional links among IPs and files.

Acknowledgements

This work is supported in part by the MAPAP SIP-2006-PP-221003 project.

References

1. V. Batagelj, A. Mrvar: Pajek – Program for Large Network Analysis. Home page: http://vlado.fmf.uni-lj.si/pub/networks/pajek/.
2. F. Aidouni, M. Latapy, C. Magnien (2008a): One week measurement on an eDonkey server, http://content.lip6.fr/latapy/edonkey/oneweek/.
3. F. Aidouni, M. Latapy, C. Magnien (2008b): Large-Scale Measurements on eDonkey Servers - data collection and management methodology, http://content.lip6.fr/latapy/edonkey/oneweek_documentation.pdf

Appendix

Appendix 1: Components in the file-IP network after units (files and IPs) with very low index of "paedophile tendency" based on the words associated with them and their neighbours have been eliminates

Appendix 2: Vertex islands in the file-IP network

Appendix 3: Vertex islands in the file-IP network with added "pedophilic" words and connections of files and IPs to these words.

Project MAPAP SIP-2006-PP-221003.

http://antipaedo.lip6.fr

Supported in part by the European Union
through the *Safer Internet plus Programme*.

http://ec.europa.eu/saferinternet