Technical report on the

Content Rating and Fake Detection System

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

Jean-Loup Guillaume, Matthieu Latapy[1], Clémence Magnien and Guillaume Valadon

Abstract

The goal of our content rating and fake detection system is to automatically detect files having paedophile or pornographic content in a P2P system. It evaluates the content of a file using user preferences, the idea being that if a given user is interested in two files, this means that these files are somewhat related. Based on this idea, we build a graph of common interest between files and extract a hierarchical community structure from it, grouping together similar files. Starting from a file list initially considered as paedophile or pornographic, this method characterises all files that are similar to the ones from the initial list, *i.e.* files with paedophile or pornographic content. In this report, we present the method we implemented and assess the quality of results using various complementary approaches.

1 Introduction

Many files in peer-to-peer systems have a pornographic or paedophile content. A first idea to identify such files is to study their names, which are supposed to describe their content. However, this is more difficult than it may seem at first glance. First, some files are *fakes*: their name differs significantly from their content. Some files may also have several names in the system, not all equally explicit, and the user may not have access to all these names. In a P2P system, there are also many situations in which no name is known for a given file. Last but not least, it is not always easy to identify correctly pornographic or paedophile file names.

The goal of the content rating system is to detect automatically which files have paedophile or pornographic content, and which files are fakes. A first version was implemented using file names [8]. Although it gave interesting results, it remained limited (to files with names, in particular), and its performances needed improvement.

This report presents a new approach for content rating and fake detection system. It evaluates the content of a file using user preferences as observed in the P2P system itself. The underlying idea is that if a given user is interested in two files, this means that these files are somewhat related. In addition, they are more related if many users are interested in these two files. Based on this idea, we build a graph (or network) of common interest between files. We then extract a hierarchical community structure from this graph, which groups together similar files.

Starting with an initial list of files known to be paedophile or pornographic, this approach makes it possible to characterise all files that are similar to the ones we start

[1]Contact author: Matthieu.Latapy@lip6.fr

from. This method has the advantage that it is possible to rate files which do not have names in our dataset, which was not possible in the previous version.

In this report we first present the method we used in full details. We then present the results obtained for paedophile and pornographic content rating and fake detection. Finally, we use two complementary validation approaches to assess the relevance of obtained results.

Our content rating and fake detection system is publicly available at `http://antipaedo.lip6.fr/Data/`. The interface allows automatic queries, so that end-user may access the evaluation of this system before downloading a given file. P2P clients may include this facility in their software. We also present the results in our web interface for data browsing, thus making it possible to assess them. Details about both accesses are presented in [7].

2 Method

This section describes the method used for computing a content rating for any file in our dataset. It is based on the observation of which users provide which files, which indicates closeness between different files. It also needs as input a list of reference files for which the rating should be maximal. Based on users preference, the method then determines if the other files are similar to the files in the list or not.

The different steps of this methodology are detailed below.

2.1 Common interest graph

Our method relies on data about user behaviours in the system. We need to know which users are interested in which files in order to build a graph of common interest between files.

In the context of this study, the method is applied to the first server measurement made in the project [1, 10]. This methodology is however very general. This data consists of a ten weeks measurement of all queries processed by a large *eDonkey* server. It contains all the queries made by users, and all the answers made by the server. There are two main types of queries:

- keyword queries;
- source queries.

Keyword queries are entered to search for files corresponding to a particular topic. For instance, a user interested in songs by the artist Madonna will make a query containing the search string `Madonna`. The server answers to such queries by providing lists of files matching the keyword(s) they contain (in our example, these files will have the string `Madonna` in their names). When a file of interest is identified, the user may then send a *source query* for this file to the server. The server answers with a list of providers for this file.

We then need to formalise the notion of a user *interest* for a file. As the data indicates which users provided which files, and which users made queries for which files, several

options exist. In particular, users can show interest for a file by providing it, or by making queries for it. We chose to take only into account the fact that a user provides a given file. Indeed, the interest expressed by the user in this case is stronger (it is easier to make queries than to provide a file, and users in general make queries for a larger number of files than they provide). The corresponding information is therefore richer, and more reliable.

We encoded this information into a *bipartite* graph $G_B = (P, F, E)$, where P is the set of observed peers, F is the set of observed files, and E is the set of links between peers and files. There exists a link between a peer p and a file f if p has been observed to provide f at least once during the measurement.

Figure 1 presents an example of such a bipartite graph. Peers are represented at the top of the drawing and identified by numbers; files are represented at the bottom and identified by letters. In this example, peer 1 provides files A, B, C, peer 2 provides files B, C, D, peer 3 provides files C, E, and peer 4 provides files D and F.

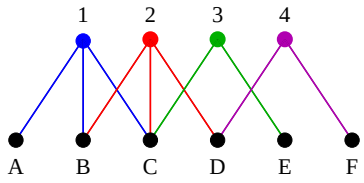

Figure 1: An example of a bipartite graph between peers (top) and files (bottom).

Note that our choice implies that some of the files we observed in the measurement are not present in G_B, and cannot be rated. However, these are the files that have been queried for, but for which no provider was observed. This means that the server never answered the corresponding queries, which indicates that these files were never actually present in the system during the measurement. As we observed a large number of queries seemingly coming from automatic crawling robots [1], we believe most of these files are not real files, and taking them into account would introduce a strong bias in the data.

We now use the bipartite graph to construct a common interest graph among files, denoted by G. This graph reflects interests of users for different files, and is built as follows: two files are linked if and only if a same peer provides both of them. In other words, there is a link between files f_1 and f_2 in G if and only if there exists a peer p such that there are links between p and f_1 and p and f_2 in G_B. This graph is called a *projection* of the bipartite graph [4].

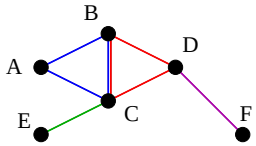

Figure 2: Common interest graph among files, obtained by the projection of the bipartite graph presented in Figure 1.

Figure 2 presents the projection of the example graph of Figure 1. Notice that all links are not equivalent in the projected graph G. For instance, the link between B and C is created by two distinct peers, 1 and 2, while other links are created by only one peer. It can also be argued that the link between A and B is weaker than the link between D and F, because the first one is created by peer 1 which provides 3 different files, while the other is created by peer 4 which provides only 2 files, thus indicating that these files are probably more similar than the ones provided by peer 1.

It is possible to take into account these differences by assigning *weights* to the links of G which reflect the importance of each link: the larger the weight, the closer (the more similar) the corresponding files are.

Inspired by [4], we considered three different types of weights on the links:

- no weight (*i.e.* all links are equivalent, as in the graph of Figure 2);
- sum weight: as a link (u, v) in the graph can be created by more than one peer, if different peers provide both u and v, the sum weight of (u, v) is the number of distinct peers that provide both u and v;
- delta weight: this definition is based on the intuition that the more files a peer provides, the less any two of these files are likely to be similar. We therefore assign to each peer a total weight of 1 to distribute equally among all links it induces. For instance, if a peer p provides 3 files, then each pair of these files receives a weight of $1/3$. If u provides 4 files, then each pair receives a weight of $1/6$. In general each pair of files receives one over $d(d-1)/2$, where d is the number of files provided by p. The delta weight of a link between files is equal to the sum of the weights given to it by all users;
- Jaccard weight: this definition is based on the idea that if different peers provide similar lists of files then these files are strongly related. Let us denote by $N(f)$ the set of peers linked to a file f. For each pair (f_1, f_2) of files, the Jaccard weight is then equal to $|N(f_1) \cap N(f_2)|/|N(f_1) \cup N(f_2)|$, *i.e.* the total number of peers who provide both files, divided by the number of peers which provide at least one of them.

2.2 Communities

The next step in our process towards rating files is to group together files that are close to each other, *i.e.* isolate groups of files between which there is a large number of strong links, whereas they are few links to files in other groups. Such groups are classical objects in complex network analysis, called *communities* or *clusters*.

The method of [2] is used to compute communities in the common interest graph G. This method has the advantage of producing a hierarchical decomposition of a graph into communities, which is crucial here. Moreover, it is the only one available which performs high-quality community detection within graphs of millions of nodes in a reasonable time.

Figure 3 presents an example of the obtained decomposition. The graph on the left contains three large communities (in light grey), one containing 3 nodes, one containing 15 nodes and the third containing 6 nodes. Two of these communities are themselves

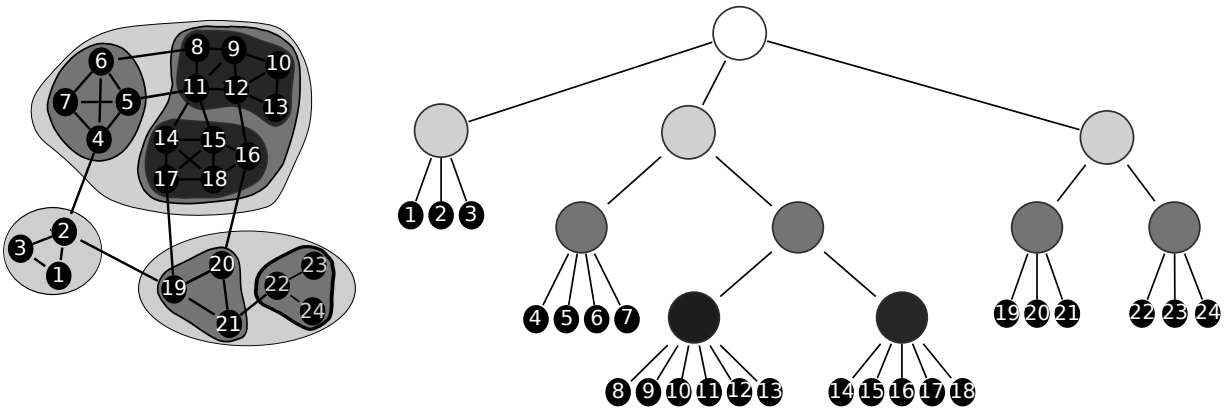

Figure 3: Example of a hierarchical decomposition of a graph into communities.

composed of sub-communities (medium grey) which in turn can be composed of sub communities (dark grey). The tree on the right describes the hierarchical decomposition of the graph (*i.e.* inclusion relations between communities): the whole graph is represented as the root (white node), at the top of the tree; its children are the three main communities (they are linked to the root and represented under it); generally, the children of a community c the sub-communities of c. The leaves of the tree are the nodes of the graph (in black). Note that some communities may be decomposed many times in sub-communities while others are not; the tree is then unbalanced.

The quality of a decomposition may be assessed by the *modularity* function, which quantifies how many links are inside communities, and how many are between files of different communities.

We computed the communities of the interest graph G, using the different definitions of weights for the links described above. The obtained community structure varied greatly depending on the weights.

The best community structure, according to the modularity function, was obtained by using *Jaccard* weights. Manual inspection of the community structures also indicated that it was more relevant (the grouping of files with names reflected better the similarities between the content they described). Therefore, we focused on the community structure obtained by using the *Jaccard* weights.

2.3 File lists

Finally, our content rating system uses the obtained hierarchical community structure to find files with paedophile or pornographic content. However, the community structure does not in itself contain any information about the nature of files, but only about which groups of files are *similar* to each other (according to user interests).

We will use this notion of similarity for our method: we give in input a list of known files to the content rating system, which uses the hierarchical community structure to infer groups of files which are *similar* to the files given in input, *i.e.* files which tend to be in the same communities as the input files.

In our context, we therefore need two file lists to provide to the content rating system: a list of files considered as paedophile, and a list of files considered as pornographic.

Since we do not have access to file contents, and no other way to construct large and relevant lists, we used file names to build initial lists of paedophile and pornographic files.

In order to build our initial list of paedophile files, we used the filter designed for automatic detection of paedophile queries developed in the project [6], which is the most accurate tool currently available for doing so. We applied this filter to all filenames observed in our dataset, and isolated files which had names detected by this filter. This list contains 12 943 files.

Note that this filter is not perfect. In particular, it has *false positives*, *i.e.* it marks as paedophile some file names which are not paedophile, like for instance:
`Kid Rock, Limp Bizkit, Korn, Eminem - Fuck Off.mp3`.
It also has some *false negatives*, *i.e.* it does not recognise as paedophile some file names which are paedophile. See the report on automatic detection of paedophile queries [6] for more details about this. However, we recall that our aim is not to classify files according to their names: our purpose is to build an example initial list which contains a large proportion of files with paedophile content, and using names is an appropriate way of doing so. We will see in the sequel (Section 4) that our method is actually able to detect files that we initially misclassified as paedophile.

We used a similar method for constructing an initial list of files with pornographic content. We started from the filter designed for matching *paedophile* queries, which contains a section dedicated to pornographic keywords used to check their use with terms related to childhood [6]. We isolated this section and used it to make a manual inspection of file names. It appeared that the general context of pornography uses a wider range of keywords than the paedophile context. Conversely, some words indicating paedophile content when they are associated to words referring to children, are not strong enough to indicate pornographic content in themselves (for instance, the word *fuck* is used in many song titles without indicating pornographic content).

After this manual inspection, we therefore added some pornographic keywords to our filter, and removed some that were not explicit enough, in order to obtain our final pornographic file name filter. We applied it to all file names in our dataset, obtaining this way an initial pornographic file list for our content rating system. Finally, we added to this list the files detected by the paedophile filter which were not detected by the pornographic filter. The final list contains 792 037 files.

2.4 Rating

Given a tree representing the hierarchical decomposition of a graph into communities, and a file list given as an input, our content rating system aims at assigning a *rating* to each file. The rating of a file will reflect its similarity (according to the community structure) to the files in the initial list.

Several possibilities exist for assigning a rating to each file. Most depend on the *ratio* of paedophile files in each community of the tree. Therefore, we computed, for each community (excluding the leaves, which correspond to single files), its ratio of files belonging to the input list vs its total number of files with a name. Note that communities which contains files without names are not assigned any ratio.

We can then assign to any file a rating equal to the ratio of the smallest community containing it (*i.e.* the first one above it in the tree), or to the ratio of the largest community containing it (*i.e.* the closest one to the root). Both choices make sense: the largest community is the one that is the final result of the community detection method, and therefore the most relevant, as assessed by the modularity function. However, these communities are often large, and a small number of files can be quickly diluted in these communities. Conversely, the smallest community that contains a node contains the nodes that are the most similar to it, and therefore its rating closely reflects the nature of this community. However, these communities are often very small and therefore not very relevant. If for instance two files from the list are grouped in a 2-nodes community, the rating of the corresponding community will be equal to 1, the maximal value, whereas it is too small to be statistically representative.

After some manual inspection of the community structure and the associated ratios, we observed that there was a high variation in the sizes of communities, as well as in their distance from the root. Therefore, the rating of each individual file must take into account the ratios of *all* communities which contain this file, from the smallest one to the root of the tree, which contains all files.

Finally, we have implemented the following rating method: for each file, we associate a rating equal to the average ratio of all communities containing this particular file.

The obtained ratings range from 0 to 1. The closer a given rating is to 1, the more similar the corresponding file is to the files of the input list.

3 Content rating

In this section, we apply the rating system described above to our dataset, using as input the lists described above. We then study the obtained results, which leads to notions of statistically relevant detection of paedophile and pornographic content.

3.1 Paedophile rating

We computed the paedophile ratings using as input list of paedophile files the one described in Section 2.3. Note that the method produces ratings for *all* files, including the ones that belong to this input list. Since we already classified these files as paedophile, we will not study their ratings in detail here. These ratings will however be useful for fake detection, see Section 4, and for the validation of our method, see Section 5.

File set	avg. rating
not paedo	0.000742363
paedo	0.0056055

Table 1: Average paedophile ratings of different types of files.

Table 1 presents the average rating for all files that do not belong to the input paedophile file list, and the average rating for the files belonging to this list (for comparison). The average rating of paedophile files is almost ten times higher than the average rating

of other files, showing that the rating system succeeds in capturing the paedophile nature of a file.

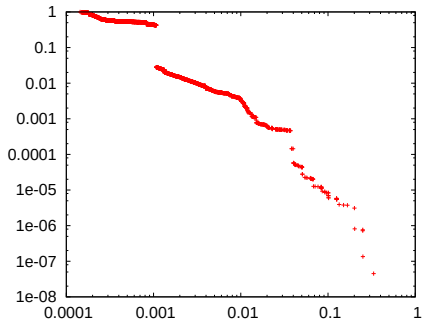

Figure 4: Complementary cumulative distribution of paedophile ratings for all files that are not initially given as paedophile.

We must now decide for each file that is not in the input list whether it has paedophile content or not, based on this rating. In order to do so, we study the complementary cumulative distribution of paedophile ratings for all files we need to classify, *i.e.* all files that are not in the paedophile list. See Figure 4. A point (x, y) in this plot means that a fraction y of the considered files has a rating greater than or equal to x.

This distribution presents two regions of high density (*i.e.* the plot decreases sharply at these points): one at $x = 0.001073$, and the other at $x = 0.0398135$. These points mean that a comparatively large fraction of the files have ratings close to or equal to these values. This corresponds to a notion of statistical *normality* in the distribution, which points out different groups of files.

Indeed, these two points divide all files into three statistically relevant groups: approximately 97.2% have a rating lower than or equal to 0.001073; 0.014% have a rating greater than or equal to $x = 0.0398135$, and the rest have intermediate ratings.

These groups induce natural categories for our content rating system. The first category contains the files with the lowest ratings, which we therefore classify as *not paedophile*. Conversely, we classify the files in the group with the highest ratings as *probably paedophile*. Finally, there is more uncertainty about the group with intermediate ratings, which represents 2.8% of the files; we classify them as *maybe paedophile*. The fractions of files in each category are summarised in Table 2.

Category	# files	Percentage
not paedophile	21 548 680	97.2%
maybe paedophile	623 059	2.8%
probably paedophile	3 196	0.014%

Table 2: Number of files classified in each paedophile category by our content rating system.

3.2 Pornographic rating

We computed the pornographic ratings using the method described above, the input list of pornographic files being the one described in Section 2.3. As for the paedophile content rating, we will not study the ratings of files from this list here, because they were already classified as pornographic.

File set	avg. rating
not porn	0.0383181
porn	0.221733

Table 3: Average pornographic ratings of different types of files

Table 3 presents the average rating for all files that do not belong to the input pornographic list. For comparison, it also presents the average rating for the files belonging to the list. The average rating of pornographic files is almost ten times higher than the average rating of other files, showing that our method succeeds in capturing the pornographic nature of files.

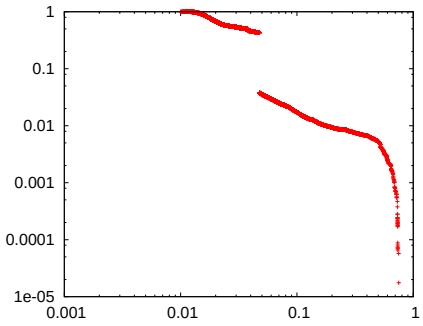

Figure 5: Complementary cumulative distribution of pornographic ratings for all files that are not initially given as pornographic.

This distribution presents two regions of high density (*i.e.* the plot decreases sharply at these points): one at $x = 0.001073$, and the other at $x = 0.0398135$. These points mean that a comparatively large fraction of the files have ratings close to or equal to these values. This corresponds to a notion of statistical *normality* in the distribution, which points out different groups of files.

Indeed, these two points divide all files into three statistically relevant groups: approximately 97.2% have a rating lower than or equal to 0.001073; 0.014% have a rating greater than or equal to $x = 0.0398135$, and the rest have intermediate ratings.

We must now decide for each file that is not in the input list whether it has pornographic content or not, based on this rating. In order to do so, we study the complementary cumulative distribution of pornographic ratings for all files we need to classify, presented in Figure 5. A point (x, y) in this plot means that a fraction y of the considered files has a rating greater than or equal to x.

This distribution presents a region of high density at $x = 0.047621$. This point means that a comparatively large fraction of the files have ratings close to or equal to this value. This corresponds to a notion of statistical *normality* in the distribution, which points out two different groups of files: approximately 96.2% of the files have a rating lower than or equal to 0.047621; and approximately 3.8% have a greater rating.

These groups induce statistically significant categories for our content rating system. The first category contains the files with the lowest ratings, which we classify as *not pornographic*. Conversely, we label the group with the highest ratings as *probably pornographic*. The fraction of files in each category are summarised in Table 4.

Category	# files	Percentage
not pornographic	20 592 213	96.2%
probably pornographic	803 629	3.8%

Table 4: Number of files classified in each pornographic category by our content rating system.

4 Fake detection

We recall that a *fake* is a file with a name that does not correspond to its content. We are interested here in the fakes for which the name or the content is of paedophile or pornographic nature. There exist other types of fake, for instance music files with a name depicting a video game, but we do not consider them here.

We may therefore observe two types of fakes:

1. files with a paedophile (resp. pornographic) name but with a non-paedophile (resp. non-pornographic) content;

2. files with a non-paedophile (resp. non-pornographic) name but with paedophile (resp. pornographic) content.

We will study these two types of fake separately, both using the ratings provided by our content rating system, see Section 3. Note that files which do not have a name cannot be fakes, therefore we do not consider them.

4.1 Paedophile fake detection

Figure 6 presents the complementary cumulative distribution of paedophile ratings, for all files that have a paedophile name. Intuitively, files with a low rating are fakes of the first category described above. Here, most files have a rating greater than or equal to 0.001073. A large fraction of them has a rating exactly equal to 0.001073, which reveals as explained above a notion of statistically relevant *normality*. Files with smaller ratings therefore naturally belong to a category of files with ratings lower than normal. This represents approximately 3.4% of the total, which we therefore classify as fakes of the first category.

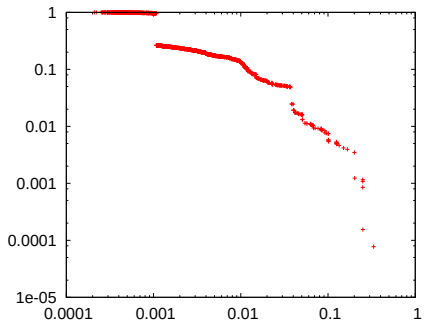

Figure 6: Complementary cumulative distribution of paedophile ratings for all files with a paedophile name.

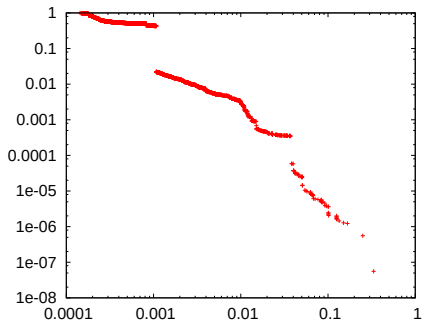

Figure 7: Complementary cumulative distribution of paedophile ratings for all files with a non-paedophile name.

Conversely, Figure 7 presents the complementary cumulative distribution of paedophile ratings, for all the files which do not have a paedophile name. Intuitively, files with a high rating are fakes of the second category. Like the previous distribution, this one presents two regions of high density that define normal and abnormal behaviours. We are interested here in files with higher ratings than normal. These are defined by the second high-density point, at $x = 0.038$. Files with a rating higher than this value represent approximately 0.0059% of the total, and we classify them as fakes of the second category.

Paedophile fake category	# files	Percentage
(1) paedophile name but non-paedophile content	443	3.4%
(2) non-paedophile name but paedophile content	1 055	0.0059%

Table 5: Number of paedophile fakes of the two categories of interest here.

Table 5 gives a summary of the number of paedophile fakes of each category.

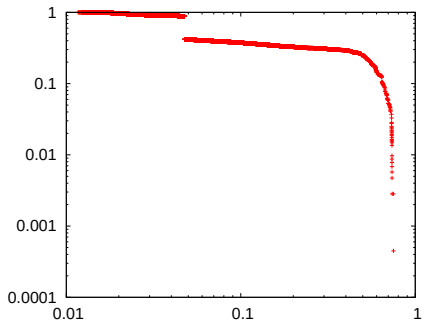

Figure 8: Complementary cumulative distribution of pornographic ratings for all files with a pornographic name.

4.2 Pornographic fake detection

Figure 8 presents the complementary cumulative distribution of pornographic ratings, for all files that have a pornographic name. Intuitively, files with a low rating are fakes of the first category. Here, most files' ratings are greater than or equal to 0.047621. A large fraction of them has a rating exactly equal to this value, which represents again a notion of statistically significant normality. Files with smaller ratings therefore have a rating lower than normal. They represent approximately 11.7% of the total, and we classify them as fakes of the first category.

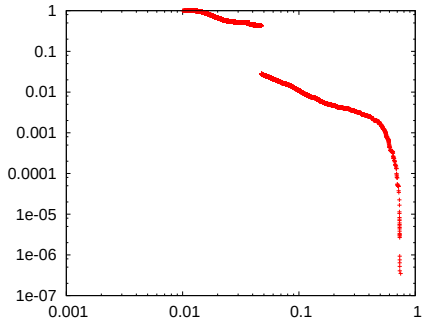

Figure 9: Complementary cumulative distribution of pornographic ratings for all files with a non-pornographic name.

Conversely, Figure 9 presents the complementary cumulative distribution of pornographic ratings, for all the files which do not have a pornographic name. Intuitively, files with a high rating are fakes of the second category. As the other distribution, this one presents a region of high density at $x = 0.047621$ that defines normal and abnormal behaviours. We are interested here in files with abnormally large ratings. These are files with rating greater than 0.047621, which represent approximately 2.8% of the total. We classify these files as fakes of the second category.

Table 6 gives a summary of the number of pornographic fakes of each category.

Fake category	# files	Percentage
(1) pornographic name but non-pornographic content	92 664	11.7%
(2) non pornographic name but pornographic content	479 255	2.8%

Table 6: Number of pornographic fakes of each category.

4.3 Files with several names

Some files are observed with several names. In most cases, these names are variations of each others (conversions between upper and lower case, translations in different languages, etc.) which describe the same content. In some (rare) cases, though, different names indicate different types of contents. These files are by definition fakes.

Dealing with such fakes, and in particular automatically detecting them, is difficult. We implemented previously a fake detection system based on the differences between the names of a same file, see [8]. Manually comparing this to the system presented here gives a good intuition of whether a given file is a fake or not, but combining them into a single fake assessment tool seems difficult. Our interface for accessing our fake detection system therefore provides evaluation provided by both systems.

5 Validation

It is essential to evaluate the relevance of the results of our content rating and fake detection system, but this is a challenging task. Indeed, we would in principle need a reliable database of known paedophile files, or pornographic files. To bypass this difficulty, we conducted two different validation procedure, described below.

We performed such evaluations for both our paedophile rating and fake detection, and our pornographic rating and fake detection. We obtained similar results in each case, so in this section we only present our validations for the paedophile content rating and fake detection results.

5.1 Structural validation

The first validation consists in studying the sensitivity of our rating method to changes in the input file list. To explore this, we created another paedophile input list, composed of a random sample of 90% of the original paedophile file list. We then computed ratings of *all* files as before, but using this new, smaller input list.

Figure 10 presents the complementary cumulative distribution of the ratings for three groups of files: all files, the files belonging to our new input list, and the 10% of files that we have removed from the initial input list. This plot evidences several important facts.

First, the distribution of ratings of all files is very similar to the one obtained previously with the full list, see Figure 4. The distribution of ratings of paedophile files is also very similar to the one previously observed, see Figure 6. This shows that globally, the ratings are not affected by small (but significant) variations in the input list, and are therefore robust.

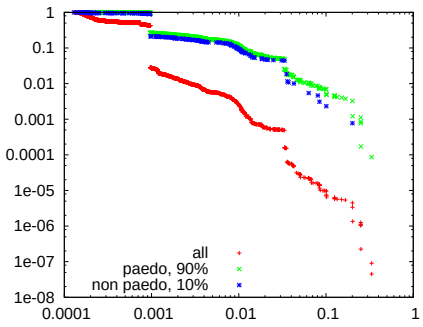

Figure 10: Complementary cumulative distributions of paedophile ratings when only 90% of files with paedophile names are taken into account for the ratings.

The plot also shows that the distribution for the 90% of paedophile files in the new input list is very similar to the one for the 10% that we removed from the initial list. This shows that our content rating system is able to capture the paedophile nature of these files, even if they are not initially listed as paedophile.

5.2 File names

Another, complementary validation technique consists in studying the names of files which are classified as paedophile, non-paedophile, and fakes, thus gaining insight on their probable content.

We studied names of the following file categories:

- 50 files with paedophile names with the highest paedophile ratings;
- 50 files with paedophile names with the lowest paedophile ratings;
- 50 files with non-paedophile names with the highest paedophile ratings;
- 50 files with non-paedophile names with the lowest paedophile ratings;
- 50 files without names with the highest paedophile ratings.

For all these categories, we studied the names that are present in our dataset. Going further, we also tried to obtain new information on these files by searching for them (identified by their *hash*) in two popular *eDonkey* search engines [3, 9]. The listing of names obtained in both cases is given in Appendix A [2].

We then inspected these names and counted the paedophile ones. Since the filter is prone to errors, we used our own expertise. File names encountered were not particularly difficult to classify, and several members of the project gave the same classification independently. We marked these file names with a '+' sign in the listing in Appendix A.

Table 7 sums up the number of files with at least one name in each category, together with the number of files with at least one paedophile name.

[2] We did not provide the names of files from the fourth category, *i.e.* files with non-paedophile names with lowest paedophile ratings, because they contained no paedophile names.

File set	# files with name	# files with paedophile name	%
paedophile names, highest ratings, (1)	50	42	84%
paedophile names, highest ratings, (2)	13	10	77%
paedophile names, lowest ratings, (1)	50	11	22%
paedophile names, lowest ratings, (2)	26	5	19%
non-paedophile names, highest ratings, (1)	50	13	26%
non-paedophile names, highest ratings, (2)	26	5	19%
non-paedophile names, lowest ratings, (1)	50	0	0%
non-paedophile names, lowest ratings, (2)	7	0	0%
no name, highest ratings, (2)	20	6	30%

Table 7: In each line: for files in a given category (first cell), the number (second cell) of file names we found (1) in our dataset, or (2) in search engines, which are of paedophile nature (third cell) and the corresponding fraction (last cell).

This leads to the following observations: first, when we study files with a paedophile name detected by the filter, the number of files with an actual paedophile name is much higher for files with the highest ratings than for files with the lowest ratings (it is approximately four times higher). The names are not a reliable indicator of the content of a file, but the higher concentration of files with paedophile names in the groups with high ratings shows that our method succeeds in creating groups with a much larger fraction of paedophile files than others.

Second, when we study files with paedophile names detected by the filter, we observe that a large fraction of the names that are not classified as paedophile by human experts are false positives of the filter. For instance, names such as:
`kid rock limp bizkit korn eminem fuck off mp3`
are detected by the filter, though they are not paedophile in themselves.

It is interesting to notice that the vast of majority of these files are assigned low paedophile ratings, and are therefore classified as fakes by our fake detection system. In fact, the ratio of files classified as fakes by the fake detection system, which is equal to 3.4%, is close to the false positive rate of the filter for paedophile string detection, see [6]. This shows that, even though the input list of paedophile files we provided to the content rating system contained some errors, the system was able to detect that these files were different from the other files from the list. This may be useful for correcting errors made by the filter.

Finally, we also studied files with non-paedophile names or with no names in our dataset. We considered the ones with highest paedophile ratings[3]. The fraction of files with a paedophile name in these sets is not much larger than the one for paedophile files with the *lowest* ratings, which might seem to be a poor result at first glance. However, two factors must be taken into account. First, fractions larger than 20% are *orders of magnitude* larger than what could be expected by sampling files at random. This is because the global fraction of paedophile files is very low. The relativity of this result is

[3]For files with no name in our dataset, the only name set we have are the recent names for these files obtained through dedicated search engines.

confirmed by the study of the names of files with no paedophile names having the *lowest* ratings (not presented here): not a single one of the file names is paedophile. The second factor is that some of the files which do not have a paedophile name are fakes. It is however impossible to know precisely the fraction of fake files among files with highest ratings without comparing this to the actual content of files.

6 Conclusion and perspectives

In this report, e presented a content rating and fake detection system based on user interest for files, as observed in P2P measurements. Our method is based on recordings of which users provide which files. Using an initial list of files known to have paedophile (resp. pornographic) content, our method is able to find files *similar* to files in the list. This has the advantage of not relying on filenames, contrary to previous versions of our content rating and fake detection system.

We presented the results we obtained in practice and evaluated them using two complementary methods. This showed that the system succeeded in grouping similar files together, and that it was robust to changes in the initial file list. Moreover, if some non-paedophile (resp. non-pornographic) files are placed in the starting list, the method generally succeeds in detecting that they are different from the other files in the list.

Several directions are possible for improving these results. First, in this version of the system, a user is identified with an IP address. However, recent findings [5] indicate that it would be more relevant to consider that an individual user is identified by an IP address *and* a specific port. Using this new definition of a user may lead to a significant improvement of the results.

Another direction would be to improve the initial list of files known as paedophile (resp. pornographic) provided as input to the system. In this work we here lists selected according to file names, which may be misleading. Despite this, our results showed that the system was able to identify most of these files. However, providing as input a list of files assessed as paedophile by law-enforcement institutions may lead to an improvement of the results.

Fakes are created by users intentionally, either to avoid detection by law-enforcement personnel, or to fool users downloading these files. Therefore, it is possible that some users create many fakes. Detecting such users would also help in improving the fake detection system.

Finally, an assessment of the results by law-enforcement institutions may lead to relevant insight on our method and hence to improvements in our method.

Acknowledgements. This work is supported in part by the MAPAP SIP-2006-PP-221003 and ANR MAPE projects.

References

[1] Frédéric Aidouni, Matthieu Latapy, and Clémence Magnien. Ten weeks in the life of an eDonkey server. In *Proceedings of HotP2P'09*, 2009.

[2] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Journal of Statstical Mechanics: Theory and Experiment*, page P10008, 2008.

[3] Figator.com. `http://figator.com/edonkey/`.

[4] Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30:31–48, 2008.

[5] Matthieu Latapy, Clémence Magnien, and Raphaël Fournier. Technical report on *Quantification of Paedophile Activity in a Large P2P system*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

[6] Matthieu Latapy, Clémence Magnien, and Raphaël Fournier. Technical report on the *Automatic Detection of Paedophile Queries*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

[7] Matthieu Latapy, Clémence Magnien, Fabien Tarissan, and Guillaume Valadon. Technical report on *Database Specification and Access*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

[8] Matthieu Latapy, Clémence magnien, and Guillaume Valadon. First report on *Database Specification and Access including Content Rating and Fake Detection system*, 2008. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

[9] Power-portal ed2k stats. `http://www.power-portal.to/ed2kstats/`.

[10] Ten weeks measurement on an edonkey server. `http://content.lip6.fr/latapy/edonkey/weeks-1.0.1/`.

A File names

A.1 Paedo names – highest ratings

A.1.1 Names in dataset (42/50)

+ 10735344 webcam 9yo karin shows cunt avi
+ 10961829 xx sexo que boa cona de adolescente porto a puta da minha mulher videos rar
110233634 young buck dj whoo kid are you a window shopper g unit radio pt 15 fuck bitches pt 2 mp3
+ 118869916 look at this porno asian amateur wife bondage lick suck fuck nude hardcore assfuck cum pussy fetish lesbian cum ddoggprn masturbate wild sado maso slave mature teen chicks slu rar
+ 12446018 hussyfan kelly model net 15yr legal very sweet complete -901958 -3517 -901959 feb -353 preview kelly -73 -1744 jpg
141608034 italia forza elezioni -202 le false opere di berlusconi in 5 anni del suo governo gba ps2 game zelig nfs fifa pes divx xxx pps
+ 16211075 15yr -801909 by -384504 messina amatoriale cel jpg
162273195 kid rock cadillac pussy mp3
+ 178292060 el videoclub en casa todas las novedades peliculas films accion aventura terror porno suspense drama amor infantiles humor url zip
+ 178292061 el videoclub en casa todas las novedades peliculas films accion aventura terror porno suspense drama amor infantiles humor url zip
+ 19160252 jennifer love fox sports naked shaved pussy upskirt twat snatch cunt qwerty sex xxx porn ddoggprn -2162866 jpg
+ 19281676 gay dad son and grandpa jpg
+ 20363096 15yr -801909 by -384504 messina amatoriale cel jpg
+ 221852683 7yo amber sucks dad s cock and gets assfucked avi
+ 227245359 taboo mother s lust kay parker is a lustful mom who sucks fucks her hung teenage son mpg
+ 228457293 hardcore sex with myself xxx lolita qwerty ddoggprn anal pussy penis suck dick vagina asshole orgy asia asian lesbo dildo jpg zip
+ 231336033 mom and daddy fucks daughter for bad behaviour jpg
+ 2318409 6y orgy leila de santos cuban bee brasil cuban teen mpeg
+ 238585396 mom and daddy fucks daughter for bad behaviour swf
+ 23924748 dad fuck daughter 08 jpg
+ 24539290 babyj sunshine 4 yo fucked she rubs cum on her pussy avi
+ 247147914 kingpass webcam karin 10yo part 4 rar
+ 249138890 kingpass newest st petersburg 02 avi -4067584 rar
254138454 kid rock limp bizkit korn eminem fuck off mp3
265087405 infantil musica para beb<C3><A9>s buenas noches 09 14 yenc 07 cancion de cuna de brahms mp3 01 12 mp3
+ 26709847 con hardcore sex with myself xxx lolita qwerty ddoggprn anal pussy penis suck dick vagina asshole orgy asia asian lesbo dildo rar
+ 2672079 fotos de la zorra de motril desnuda venus 15 anos by antispam sex porno puta warra cria ile zip

+ 27415929 inga ass 5yo mpeg
27618172 alex kid feat lisette alea dont hide it josh wink acid pussy interpretation mp3
+ 3094845 new ptsc valery 9yo gostosinha -356045 rar
+ 3094857 new ptsc valery 9yo gostosinha -90031 rar
+ 31836813 gerl 4yo sperm jpg
+ 36328662 mafiasex ru lolita soft s8 03 jpg
+ 39163536 britney spears 14y old amateur movie of her butt in jeans candid ass avi
+ 45933391 webcam msn 14 yo no fake avi
+ 4606518 mom and daughter sex education avi
+ 46841971 lot of videos webcam by -288566 arab feet young 15 yo -646 yo -206 you voyeur string thong txt
47847131 film xxx ecole d adulte de cours d <C3><A9>ducation sexuel fr vbn pas besoin de licence interdit au mineur zip
+ 51330119 t -1569940 taboo mother s lust kay parker is a lustful mom who sucks fucks her hung teenage son mpg
+ 53686295 mafiasex ru lolita soft s2 03 jpg
+ 57022726 asiatic chinese video webcam netmeeting msn tits ass string thong 15 yo rar
+ 57022726 msn webcam net asiatic chinese video webcam netmeeting msn tits ass string thong 15 yo rar
5803297 simple plan 10 i m just a kid mtv hard rock live mp3
6119550 english kids educ pc game caillou s birthday party win98 2 6 yrs rar
+ 63843216 dad fuck daughter 12 jpg
+ 79880596 webcam msn 14 yo no fake msn webcam net rar
+ 8598972 inzest mom fuck her son -2038 jpg
+ 88370658 secret high school amateur hand job cumshot 11 sec -1363079 nice japanese porno xxx mature mom teach son s friends about sex mpg rar
+ 88370659 secret high school amateur hand job cumshot 11 sec -1363079 nice japanese porno xxx mature mom teach son s friends about sex mpg zip
+ 88370660 secret high school amateur hand job cumshot 11 sec -1363079 nice japanese porno xxx mature mom teach son s friends about sex mpg rar
+ 92272394 amatoriale mia cognata sandra a 15 anni e la sua bella fica jpg

A.1.2 Recent names (10/13)

```
f7cdda8d434df2760ee843c7373dfb9e World.Series.Of.Poker.2005.E32.PDTV.XviD-TBS.mpeg
+ f7cdda8d434df2760ee843c7373dfb9e XXX - Wonderful Teen Sex - Leila De Santos
- Cuban Bee - Brasil & Cuban Teen - High Quality - Agt2 Sc02.mpeg
+ fca1a92a0d651f7e083e324cc04a7d2f Fotos.de.la.zorra.de.motril.desnuda.VeNusˆ
ˆ.15.a<C3><B1>os.by.  antispam.-.Sex.Porno.Puta.Warra.Cria.ilegal.underage.zip
+ 7f65df682400e92079d35c655b343944 !!!NEW - PTSC - Valery 9Yo - Gostosinha
(G02).rar
+ 01018ab20d94f92daf7ff00a906f2cc6 !!!NEW - PTSC - Valery 9Yo - Gostosinha
(G13).rar
06643d219c510e9a3d37406c862b97cb (English Kids Educ)(PC Game) Caillou's Birthday
Party (Win98) (2 - 6 Yrs).rar
+ be5bf157f643c03283b5ae751a5e0b0f ((Hussyfan)) Kelly-model.net 15yr legal
very sweet complete-gal01-59+1wmv (Feb.2004).preview.kelly-29-24.jpg
+ d3b310740a3e7be8cee378c4861e819a Babyj - Sunshine - 04yo Fucked She Rubs
Cum On Her Pussy.avi
+ 2951180a3a2cc802a27ed83cdb2b7af1 5Yo-Inga-Ass.mpeg
+ 2951180a3a2cc802a27ed83cdb2b7af1 5Yo-Inga-Ass kleuterkutje.mpeg
e13606b0aca09d7a673a6e3703c072e7 Alex Kid feat.  Lisette Alea - Dont Hide It
(Josh Wink Acid Pussy Interpretation).mp3
+ 0b85e480af5b8d1c1028d6dcac6f7fee Webcam msn, 14 yo, no fake, avi
+ 0b85e480af5b8d1c1028d6dcac6f7fee Webcam Msn, 14 Yo, No Fake, Msn - Webcam.avi
+ 1529b5d659baca0850f5652e44cb7006 Pthc - 7Yo Amber Sucks Dad's Cock And Gets
Assfucked.avi
+ 1529b5d659baca0850f5652e44cb7006 7Yo Amber Sucks Dad's Cock And Gets Assfucked.avi
+ 1529b5d659baca0850f5652e44cb7006 Pthc - 7Yo Amber Sucks Dad's Cock And Gets
Assfucked.avi
+ 1529b5d659baca0850f5652e44cb7006 - (SDPA) Amber - oral - anal - juegos -
9 a<C3><B1>os.avi
+ bd624b11ef6322332aae0e680e5cb909 (((KINGPASS))) Webcam - Karin 10yo Part
4.rar
d09512b8b40ffaa746e194e85d020832 Relajante Infantil Musica para beb<C3><A9>s
- Buenas Noches - muy buena.mp3
d09512b8b40ffaa746e194e85d020832 Infantil Musica Para Bebes - Buenas Noches
[09 14] <C2><B7> Yenc 07 Cancion De Cuna De Brahms Mp3 (01 12).mp3
```

A.2 Paedo names – lowest ratings

A.2.1 Names in dataset (11/50)

100001269 kid loco a grand love story she s my lover a song for r mp3
+ 1021378 7yo sucks dick avi
+ 1021378 babyj sweet sucks dick mpg
+ 1021378 hussyfan my 3yo daughter suck my dick mpg
1021378 image converter 2 plus avi rar
+ 105365690 09 nofx fuck da kids mp3
119701241 10 yes my darling daughter mp3
12104456 nino d angelo sti 15 anni mp3
12897780 eric pridz i just like to call you my bitch mp3 kid kenobi and pocket remix mp3
139307930 kid loco she s my lover mp3
+ 1449966 eurololita sandra model pissen pee animal shit keys -60626 dark elf hussyfan lolitaguy rar
148876301 -361 sexy sadie what have you done 10 years of singles -353 emg m3u
1599442 eric pridz i just like to call you my bitch mp3 kid kenobi and pocket remix mp3
1599442 mylo drop the pressure kid kenobi and pocket remix mp3
16767191 italia forza elezioni -202 le false opere di berlusconi in 5 anni del suo governo gba ps2 game zelig nfs fifa pes divx xxx pps
171336009 fuck eminem fuck kid rock txt
172582056 arestra peli porno chica francesa private castings no para menores porno duro mpg
174517524 kid loco grand love story -2040 07 she s my lover a song for r mp3
+ 17455107 sexo meninas de 14 e 15 anos se amando mpeg
190121669 genesis no son of mine e 8y -279103 kar
2048223 -1095767 2 mildred reis dvdrip xvid 1 -157 3 by kaolho avi
2048223 brasileirinhas kid bengala 2 mildred reis dvdrip xvid 1 -157 3 by kaolho avi
2048223 kid bengala 2 avi
2048223 kid bengala 2 mildred reis avi
2048223 xxx kid bengala 2 dvdrip xvid avi
+ 2129316 xxx elisa -255180 nuda sul letto del motel prima di scopare 14 anni troia mia ragazza -95108 per scambio con coppie giovani -4356 -95109 jpg
217416788 kid rock american bad ass mp3
+ 2319901 mafiasex ru ren kids hard -4160787 porn jpg
23204859 pc game ita kids la fabbrica dei -890997 4 7 anni iso
2373727 kenny wayne sheperd the place you re in 03 spank featuring kid rock mp3
2373727 kenny wayne shepherd the place you re in 03 spank featuring kid rock mp3
25528316 -458364 kar
25528316 texas -458364 mid
25528316 texas summer s son 4 y 5 kar

25528316 texas summer s son kar
25528316 texas summer s son mid
259488997 05 yes sir that s my baby mp3
29832240 10 7 year bitch dead men don t rape mp3
29832240 10 dead men don t rape mp3
32776909 nashville pussy the kids are back mp3
+ 33724337 sex spanish sex menores nudist beach family kids jpg
33724337 slipknot sexy ass goth cheerleader jpg
33739324 kid rock limp bizkit korn eminem fuck off mp3
33739324 korn feat limb bizkit eminem fuck off mp3
33739324 limp bizkit kid rock korn eminem fuck off mp3
33820581 05 anorexia nervosa retrouver son etat initial -279780 avant quil ne soit trop tard vic mp3
34237004 atomix house -612 with david guetta benny benassi laurent wolf da fresh kid creme laurent konrad by dj thibaut fucking mix party 10 mp3
34237004 atomix house -612 with david guetta benny benassi laurent wolf da fresh kid creme laurent konrad mp3
42332591 nofx 15 fuck the kids ii mp3
42573974 07 ya no -339597 al son de los tambores mp3
43151394 narada guitar 15 years of collected works -3932 lover s promise david arkenstone mp3
46966691 santiago y -302071 augusto lolita cantares y juegos de las ninas -10362 txt
51424043 kid rock fuck off mp3
51920425 02 kid -507246 times hard rmx r2r mp3
53482916 nino d angelo lolita mp3
5792253 03 nino d angelo lolita wma
59212396 07 yes sir it s my s o n s<C3><AD> se<C3><B1>or es mi son mp3
62806022 seriales para panda titanium -202 son 5 y funcionan doc
+ 63179423 sexo anal adolescente mpeg
63921628 bob bigcock britney you suck limpbizkit korn eminem metallica slipknot kid rock bsb n sync mp3 685859 du jour au lendemain french dvdrip xvid lost ugm avi
+ 685859 porno adolescente -408663 avi
71855260 american bad ass 02 kid rock the history of rock uk r b 320kbps mp3
+ 719077 japan lolita yuka ptsc zip
81608661 kid creole the coconuts the sex of it extended remix mp3
84651206 04 yola da great fuck yall mp3
88375670 kid rock devil without a cause 12 fuck off mp3
92990570 whoo kid ft snoop dogg and prodigy whip yo ass extended mp3
+ 9742840 look at this porno asian amateur wife bondage lick suck fuck nude hardcore assfuck cum pussy fetish lesbian cum ddoggprn masturbate wild sado maso mpg

A.2.2 Recent names (5/26)

```
aa58735db289eda2f813ab5e5995a68a Aktuell Rapport Video Production - Lustgarden
- Sex, Porno, Girls.avi
aa58735db289eda2f813ab5e5995a68a ice age 3.avi
aa58735db289eda2f813ab5e5995a68a Oceans Of The World.avi
aa58735db289eda2f813ab5e5995a68a Schweden Porno.avi
+ aa58735db289eda2f813ab5e5995a68a Porno Adolescente Suedoise- Sweden Teen
Porno - IMAX porn - Oceans Of The World.avi
+ aa58735db289eda2f813ab5e5995a68a Porno Adolescente Su<C3><A9>doise.1.avi
aa58735db289eda2f813ab5e5995a68a Aktuell Rapport Video Production - Lustgarden
- Sex, Porno, Girls.avi
+ 626421b7b7f96fe66f7922686e7c4357 [General] [Japan Lolita] Yuka - Marchen
Story.zip
+ 626421b7b7f96fe66f7922686e7c4357 (Japan Lolita) Yuka (Ptsc).zip
+ 24a8e71201cda5ec14c3f19c71693f21 (Hussyfan) (pthc) (r@ygold) (babyshivid)
Babyj sweet 7yo sucks dick.mpg
+ 24a8e71201cda5ec14c3f19c71693f21 (Hussyfan) (pthc) (r@ygold) (babyshivid)
Babyj sweet 7yo sucks dick aka pedofilia.avi
+ 24a8e71201cda5ec14c3f19c71693f21 Olaf Framke Collection V0001 - Gay Boys
Anal Blowjob Hardcore Sex Pedo Lolita Rape Bondage Hentai Fkk Kiddy Preteen
Underage PTHC.avi
96762503129a19d93b99978478c11faf Eric Pridz - I Just Like To Call You My Bitch
Mp3 (Kid Kenobi And Pocket Remix).mp3
96762503129a19d93b99978478c11faf MYLO - DROP THE PRESSURE (KID KENOBI AND POCKET
REMIX)- TRONIK.MP3
96762503129a19d93b99978478c11faf Mylo - Drop the Pressure (Kid Kenobi and Pocket
Remix).mp3
96762503129a19d93b99978478c11faf 06 mylo - drop the pressure (kid kenobi and
pocket remix)- tronik.mp3
ea639765e4ad00545894e9a58e91bc8c Brasileirinhas - Kid Bengala 2 & Mildred Reis
200806.avi
ea639765e4ad00545894e9a58e91bc8c KidBengala.2.&.Mildred.Reis.DVDRip.Xvid.1.0.3.by.Kao
(bronhaman.com).avi
+ 8e134b01c30f14fadcc4a9ee04645893 <non-ascii characters> XXX Lolitas Kid Porno
<non-ascii characters> (83).jpg
+ 8e134b01c30f14fadcc4a9ee04645893 MafiaSex.Ru Children Kids Hard 000183 ChildPorn.jp
8e134b01c30f14fadcc4a9ee04645893 hldkfj (74).jpg
dbe43782f106dcbebc7fb4b0ed8b58dc Spank - Kenny Wayne Shepherd (featuring Kid
Rock).mp3
dbe43782f106dcbebc7fb4b0ed8b58dc 03 Spank (Featuring Kid Rock).mp3
7df3411e477f5c30e3b1021ce482cc3c 03 - nino d'angelo - lolita.wma
+ c2804f6b15972ea2a86341b35dc41100 Look At This Porno Asian Amateur Wife Bondage
Lick Suck Fuck Nude Hardcore Assfuck Cum Pussy Fetish Lesbian Cum Ddoggprn
Masturbate Wild Sado Maso.mpg
c827059301a7fd24388e2de5c518443e [Mp3] NINO D'angelo - Sti 15 anni'.mp3
```

c827059301a7fd24388e2de5c518443e Nino D'angelo - Sti 15 anni'.mp3
c827059301a7fd24388e2de5c518443e piccola mamma.mp3
4ba54f83e7f8bd97cbaa1853d11169e5 Eric Pridz - I Just Like To Call You My Bitch.mp3
(Kid Kenobi and Pocket Remix).mp3
54fafb96f92d0988e1e3f3d8af33046c Crash Bandicoot Fusion.gba
4224ba98f291da1c3a60e2759c7433d7 Texas - Summer's son.mid
4224ba98f291da1c3a60e2759c7433d7 Texas - Summer's Son (4 y 5).kar
e4ac071b094cf03704f3a883be6bb9f6 Nashville Pussy - The kids are back.mp3
e4ac071b094cf03704f3a883be6bb9f6 03 - Nashville Pussy - The kids are back -
Twisted Sister.mp3
e4ac071b094cf03704f3a883be6bb9f6 03 - Nashville Pussy - The kids are back.mp3
31142d4a1c2a069a3cab88ec6661f318 sexy ass goth cheerleader.jpg
2e4121340629d86384a6fa787c668064 Kid Rock, Limp Bizkit, Korn, Eminem - Fuck
Off.mp3
2e4121340629d86384a6fa787c668064 kid rock, limp bizkit, korn, eminem - cocky
- fuck off.mp3
2e4121340629d86384a6fa787c668064 Eminem & Korn & Limp Bizkit & Kid Rock - Fuck
Off.mp3
2e4121340629d86384a6fa787c668064 Limp Bizkit Kid Rock Korn Eminem - F.mp3
e9291da62c41de7bcdbeaa79c2ecfe64 Atomix House 2003 (With David Guetta,Benny
Benassi,Laurent Wolf,Da Fresh,Kid Creme,Laurent Konrad) By Dj Thibaut Fucking
Mix Party -10.mp3
e9291da62c41de7bcdbeaa79c2ecfe64 Atomix House 2003 (with David Guetta, Benny
Benassi, Laurent Wolf, Da Fresh, Kid Creme, Laurent Konrad..mp3
e9291da62c41de7bcdbeaa79c2ecfe64 Atomix House 2003 (with David Guetta, Benny
Benassi, Laurent Wolf, Da Fresh, Kid Creme, Laurent Konrad...10.mp3
3c03936ed21a8ce19f76dfb09d31e411 NOFX - 15.Fuck The Kids II.mp3
3c03936ed21a8ce19f76dfb09d31e411 15 - Fuck The Kids II.mp3
86a5ad652d7f36ca6b3b06ef10e20e76 07 - Ya no danzo al son de los tambores.mp3
0044c70cc600e0180b1838eca75dd04e Santiago y Gadea, Augusto - Lolita cantares
y juegos de las ninas 1910.txt
3435c90a655398e4f9ff258dc3382fee (Varsting Riddim) - - kid kurrupt - times
hard rmx - r2r.mp3
e16a68d7e81dde2d8c8ee2b790d3474f 07 Yes Sir, It's My 'S - O - N' (Si Senor,
Es Mi Son).mp3
e16a68d7e81dde2d8c8ee2b790d3474f GLORIA ESTEFAN Yes Sir, It's My 'S - O - N'
(Si Senor, Es Mi Son).mp3
849c974520fef295a60d8ac88ac7619b Whoo Kid vs Snoop Dogg vs Prodigy - Whip Yo
Ass.mp3
849c974520fef295a60d8ac88ac7619b DJ Whoo Kid ft. Snoop Dogg And Prodigy -
Whip Yo Ass (Extended).mp3
18b00443ed15310cd9498b33576a9dc5 Genesis - No Son Of Mine.mid
8a87420fd48b0172b0ab8b5db694136c Kid Rock - American Bad Ass (Uncensored Version).mp3
8a87420fd48b0172b0ab8b5db694136c Kid Rock - American bad ass.mp3
f306bc16ec4f16b2227aa35b0486bf04 05 Yes, Sir, That's My Baby.mp3

A.3 Non-paedo names – highest ratings

A.3.1 Names in dataset (13/50)

```
100399072 07 snow patrol run mp3
+ 1006527 rika nishimura and -308509 -15157 -96938 -4407944 mpg
109876066 02 chamillionaire in the trunk mp3
115905123 cycling -1440 tour de france stage 12 -108181 sur -84843 le cap d
agde -4983411 lemond wmv
11662629 -678143 12 yr mpg
117391897 battlefield -28584 keygen for -549488 tested crack serial rar
12415330 kiss do you remember rock n roll radio mp3
12415337 gregorians chants rock ballads 09 tears in heaven mp3
145872426 wwe divas -202 lingerie special torrent
+ 148826944 15yr jpg
149019120 elisa feat ligabue gli ostacoli del cuore wma
152395209 arestra jovencitas aprendiendo a follar avi
+ 170229196 qsh flashing -9994 corinne naakt nackt prive privat -219667 jpg
185803085 swordfish -5021 dominion gowenna sub
19612582 exitos instrumentales arpas y flautas de los andes 04 los indios paraguayos
-615577 mp3
+ 21227732 rika nishimura 12yo -308509 12yo 1 mpg
+ 21286334 2 qsh c -25557 julia naakt nackt prive privat -138584 my brothers
hot wife jpg
246105201 mom sue txt
247026156 xque compilation -1081 in live rar
259984731 jovanotti lorenzo -395 capo horn 07 un raggio di sole mp3
260949771 die toten hosen zehn kleine j<C3><A4>germeister mp3
3014809 -114479 do me lil mpeg
+ 32195947 webcam msn young -646 yo 15 min no fake avi
33715393 msn webcam net webcam beurette arab -3747 grosse chienne ass fucking
fist video msn -646 yo txt
33715393 webcam beurette arab -3747 grosse chienne ass fucking fist video msn
-646 yo txt
+ 33729755 lsm duch -10004 -20449 jpg
+ 3440067 video angels lsm ls magazine 07 01 mpg
35430265 msn webcam net videos msn webcam young arab tits feet foot fetish
string thong rar
35430265 videos msn webcam young arab tits feet foot fetish string thong rar
= 37207408 qsh c -9994 trista naakt nackt prive privat -4399127 jpg
4278422 bisexual shemale menage a trois avi
4341256 -247299 -79 and nasty 15 mpg
+ 48270067 inzest mutter tochter blasen sohn jpg
+ 49654027 lolitaguy ls land issue 06 little pirates sets -85 -91 -173416 pics
lsp preview jpg
5038550 erase una vez los inventores 03 heron de alejandria avi
51888346 best of trance 3 mixed by -344337 push jan johnston tiesto ferry corsten
```

lange signum hemstock jennings astral projection dumonde mp3
+ 6425455 lolitaguy home lolita series hl -19041 complete rar
64271703 ragnarok battle offline nocd patch crack serial zip
67575600 simone cristicchi studentessa universitaria mp3
7067191 marilyn manson cake sodomy live in germany mpg
76747305 msn webcam video webcam msn exhib exhibition sexy seins nus coquine amatrice cochone qui suce sex god playa nudista voyeur nude sex rar
76753873 meuf du -3747 sur msn on voie ces seins et sa chatte -112519 -1135985 -683020 blanc -302494 -63827 aulnay sous bois -850370 sous bois noisy le sec arab en chaleur rar
76754927 voyeur upskirt msn webcam msn webcam no fake very good video arab beurette big tits feet foot rar
80051447 berurier noir viva bertaga -1489 macadam circus mp3
80220497 -219878 jpg
816018 pregnant -3313071 mmmf japanese brunette black socks strips in docter s office sucks wanks to hand cum shot uses 2 way vibrator masterlive wmv
+ 85057356 webcam msn young -646 yo 15 min no fake msn webcam net rar
+ 85057357 video msn webcam 15 -646 -206 yo no fake arab webcam big tits on msn rar
88493766 smallville simple plan when i m with you mp3
88906459 wow gospel -1675 d1 t01 when my season comes bishop t d jakes the potter s house mass choir mp3
9336533 tb kdv 4 mpg
9882221 otra -310734 que ha perdido el movil o la camara de fotos xd zip

A.3.2 Recent names (6/17)

```
fc8a75bb2283832fb8b41a1dadd58de3 Pregnant-393L-Mmmf Japanese Brunette, Black
Socks, Strips In Docter's Office, Sucks, Wanks To Hand Cum Shot, Uses 2-Way
Vibrator - Masterlive.wmv
+ 05d018ba63f0e61d262ab8e6e13af66e Rika.Nishimura.and.Kayoko.(530,.850,.404bytes).mpg
+ 138d15ca28d920568c059a40653bed7e 4Yr Do Me Lil.mpeg
ac0323062d76451ee17514be8fac348c Kamiko (18 And Nasty 15).mpg
6a2433b25cffa61ffaa702565ff67f91 Erase una vez...Los inventores - 03 - Heron
de Alejandria.avi
e58b58722bb6082a0c77f44431025c8c Marilyn Manson - Cake & Sodomy [(Live In Germany].mp
+ 6f8ccc89846c0b1214e90da58da90569 TB kdv 4.mpg
1fad169a812a43916f171d1a316b49ee Exitos Instrumentales - Arpas Y Flautas De
Los Andes - 04 - Los Indios Paraguayos-Suky.mp3
+ ba164001fad41b8acb80ddba07406951 Rika Nishimura.12yo & kayoko 12yo - 1.mpg
0bbcce2c231b8c9b0f1c474a65f6bac4 go to for other videos.txt
+ 20e53be5fa830c76dc23988b458bb0d4 Inzest -Mutter & Tochter Blasen Sohn.jpg
99fd9f3842802531c4b68898d3aad1e1 best of trance 3 mixed by Audionaut (push
jan johnston tiesto ferry corsten lange signum hemstock&jennings astral projection
dumonde).mp3
519478837b0922cab4c6100f835020b5 WOW Gospel 2002 - D1 T01 - When My Season
Comes (Bishop T.D. Jakes & The Potter's House Mass Choir).mp3
4b2022f981caca0b991f2c228923ca4a 15hx.jpg
+ def8db530011f02fee07fbca11c675c3 QSH Flashing # 021 Corinne 3731.jpg
1f9dce770c27ea97af6fb5c3f83823e3 MOM&SUE.TXT
43811f524e8a1e29bc5e2a032fc891b6 Xque compilation 2007 in live.rar
```

A.4 No name – highest ratings

A.4.1 Recent names (6/20)

```
+ ce649fc6af7af8c588e26b57b26e5f02 Babyj 27 (5Yo Preteen Lolita) Strips - Self
Fingering - Fucked Girl Nobull Right On-Just About Everything And Good Quality
Too(Pthc Kiddy Pedo).avi
+ ce649fc6af7af8c588e26b57b26e5f02 Babyj.avi
fb4ca345300bad7e89b13af06861a2e3 Cassie - Just Friends.mp3
a7dad8aed225ef67e0094bf377dd4ddd <non-ascii characters> Kawanariko.mpg
01e68146beb5d21a7f24c840c76097c3 Rurouni Kenshin.  Meiji Kenkaku Romantan -
05 - Sakaba versus Zanba!  Fight the desperate fight - [TK](9efbc28a)(dub.sub
jp.de).avi
01e68146beb5d21a7f24c840c76097c3 Rurouni.Kenshin 05 (ger sub).avi
+ 9e08adc2c745e48aec9a819a66316e37 xxx (Hussyfan) Pthc New!!  Old Man 58Yr
Fuck A Young Girl Privat .avi
9e08adc2c745e48aec9a819a66316e37 Amatoriale - Anziano scopa con ragazza.avi
861f5f1f1be3e88664149290574ff7ca 1 Pondo TV Hiroko Kobayashi.avi
861f5f1f1be3e88664149290574ff7ca Yuzuki Anri <non-ascii characters>.avi
861f5f1f1be3e88664149290574ff7ca 1pondo147 - <non-ascii characters> Kobayashi.Hiroko.
+ 076801cece2456c381c3efaa3dc92db5 Pedo - Vicky 7 - Bj.avi
3f4b8c3a4959bca4917e8a1270600c1c darbouka (orientale) danse ventre - Darbuka
ikinci solo.mp3
8ea144f1b12b873a44e2935511e6f70a Japanese Busty Drunk Girl Porno.mpeg
+ a6f3fcb57cb50517ad502237414b0f7c Rebuilt <non-ascii characters> Rika Nishimura.zip
2c5c7dc422404a1dfe2bc1be03d8aa86 Elfen Lied - 01 - Chance Meeting - Begegnung
- [SL](55498776)(dub.sub jap.ger)[AniDB].avi
2c5c7dc422404a1dfe2bc1be03d8aa86 Elfen Lied -.01.(German Sub) [SL] avi
2c5c7dc422404a1dfe2bc1be03d8aa86 elfen.lied.-.01.german.sub.by hardwareguru.eu.avi
ac5ed27c819a0e6b70cc91ce7befae7c <non-ascii characters> Bandidas.2006.CD1 [MZfans.com
Mini Player].avi
ee944780e4fb2d2571027c4c39f26bac My Cock 3 cumshot mein schwanz sperm selfpics
(DickerSchwanz@kazaa) Nude FKK Dick Penis Pimmel Schwanzlutscher BI Wichser.jpg
9af62a6b0dc277a578f3e6501155ad33 uptosky.avi
f6984e424545ff76f7876e64523a431a Curro Savoy.La nuit du siffleur.02.La mauvaise
r<C3><A9>putation (G.mp3
+ 3b348217479fb99036f49e7773ecafa5 <non-ascii characters> Yuko Ogura <non-ascii
characters>.jpg
+ 3b348217479fb99036f49e7773ecafa5 Yuko Ogura (enl notepad).jpg
7f6b234e3f2e7f9a8c7d49720efa2aeb An-2 Flight Operation Manual (russian).pdf
7f6b234e3f2e7f9a8c7d49720efa2aeb Antonov An - 2 - Handbuch.pdf
7f6b234e3f2e7f9a8c7d49720efa2aeb RLE An - 2.pdf
1dff7074f17f072954ce096a775e6886 anna - kournikova - 03.mpg
+ cdffef0dcc438b37669de4145d5f662a [KTV]<non-ascii characters>.mpg
68b8b893cc34089545a78693e0f1e141 anna - kournikova - 02.mpg
```

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