Measurement and Analysis of P2P Activity Against Paedophile Content

http://antipaedo.lip6.fr/

Midterm General Public Report
– September 2008 –

Coordinator: **Matthieu Latapy**
LIP6 – CNRS and Université Pierre et Marie Curie
Matthieu.Latapy@lip6.fr

Abstract

This document provides a general-public synthetic view of the actions conducted within the *Measurement and Analysis of P2P Activity Against Paedophile Content* project and the findings derived from them. The P2P activity measurement methods and the obtained datasets are presented, as well as the results derived from keywords analysis, and a first approach for content rating and fake detection. References to relevant documentation and online material are also provided for a more detailed and technical information.

1 Introduction.

Despite the fact that dozens of millions of users are involved in P2P networks, very little is nowadays known regarding what really occurs in these networks [6]. In particular, the longstanding claim that much paedophile content is exchanged through P2P networks raises the question of the nature and extent of these exchanges, and of their control by law enforcement institutions.

In order to help addressing these issues, the *Measurement and Analysis of P2P Activity Against Paedophile Content* project has three main goals:

- designing methods and tools to protect peer-to-peer (P2P) users from harmful content;
- providing accurate information on paedophile exchanges occurring in these systems (for both general public and specialists);
- helping law enforcement institutions and NGOs to fight paedophile exchanges.

In order to achieve these goals, rigorous large-scale measurements and analysis of peer-to-peer exchanges have been conducted, at a much larger scale than what has been done so far.

The project's consortium gathers a multi-disciplinary set of European research institutions and NGO, and is funded by the European Commission and national agencies. The project is conducted in close collaboration with law enforcement institutions, which help defining relevant priorities and assessing obtained results. For a full description of the project, its rationale and goals, see [5].

This report aims at presenting our actions and findings to the general public, while this project reaches the end of its first year. In the following, each section focuses on a specific contribution. A summary is also provided at the end of each section.

2 Measurements and datasets.

The *eDonkey* network relies on a set of servers to which clients (peers) send queries. The basic communication scheme between a peer and a server consists in four steps: (1) the peer sends a keyword-based query which describes the content it is interested in; (2) the server answers by sending a list of files matching the query (more precisely: file identifiers, file names, and other descriptive elements); (3) the peer chooses some files in this list and asks the server for a list of providers; and (4) the server sends a list of providers for these files. Afterwards, the peer can directly contact the providers to get the files.

Several approaches are possible to observe the activity in this network and the following ones have been developed in the project:

- **Measurement at server level.** A capture program placed on a server registers the queries it receives and the answers it sends[1]. This way, *all* queries managed by this server are captured. Such a measurement has been conducted on a large server during a period of 10 weeks, leading to the observation of almost 9 billion messages, involving 89 million peers and 274 million files [1].

- **Measurement by client sending queries.** A client program sends queries to servers based on a set of predefined keywords to monitor[2]. This may be repeated periodically during long periods to obtain more data. Several such measurements were conducted, including a one-month long capture in which the client was connected to more than 100 servers and sent a query on 4 keywords every 2 hours. 200 000 files with filenames and 685 000 peers were observed.

- **Measurement by honeypots.** A client program advertises some files of interest (by declaring to servers that it owns these files) and then registers the queries it

[1] This kind of measurement is similar to the measurements conducted by Web search engines like *Google*, which record the queries sent by users and the answers they obtained.

[2] This kind of measurement is similar to the collection of Web data obtained by sending queries to a Web search engine and then recording its answers.

receives from other peers [3]. Several such measurements were conducted, including a one-month long one in which the client advertised 32 files and observed 24 649 peers [2].

Let us insist on the fact that, for privacy protection concerns, all the data collected within this project is strongly anonymised *during* the measurement: no personal information (in particular IP addresses) is stored at any time. See [1, 2] for details.

Notice also that the collected data does not give any view of the actual file exchanges: only *queries* and the corresponding answers have been recorded. This information is very rich, as it captures both user behaviors and the kind of files exchanged. No such information could be collected at this scale by observing actual exchanges.

Finally, the obtained datasets give complementary views of the activity in *eDonkey*: server measurements show all the activity but on one server only; client measurements focus on specific keywords or files, but may capture most of the activity concerning them. The obtained datasets are orders of magnitude larger than previously available ones. One contribution of the project consists in the public provision of these fully anonymised datasets to the research community.

In addition to the raw data, a Web interface has been developed to browse this data and get a more precise insight. Indeed, the raw data consists in a series of hundreds of millions of recorded queries, which contain very rich, but not directly available, information. For instance, one may wonder how many files (and which) a given peer has provided or downloaded; how many peers (and which) downloaded or provided a given file; which queries a given peer sent to the system; which names are associated to a given file; etc.

All this information was precomputed and a web-based interface to the results was implemented [4], see [3] for details. In this interface, one may enter a *fid* (file identifier) or *cid* (client identifier) and then obtain all available information regarding the corresponding file or peer. Moreover, as lists of peers and files are included in this information (for instance the list of peers providing a given file), one may browse the data by clicking on the corresponding data (like the *cid*s). This gives a convenient way to explore the dataset and develop an intuition on its content. Further information (like content rating and fake detection facilities, see Section 4) will be added to this interface as the project progresses.

[3] This kind of measurement is similar to the creation of a Web page and then the recording of the accesses to this page, which is usually done under the form of server logs.

[4] `http://www.antipaedo.lip6.fr/Data/`

Summary of contributions regarding measurements and datasets.
Three different and complementary approaches have been followed to observe the activity in *eDonkey* systems continuously during long periods. They provide information on billions of messages exchanged in the system, involving dozens of millions of peers and hundreds of millions of files. This is much larger than previously available measurements. Anonymised data is publicly available for research use, together with a Web interface to browse them. This interface also provides higher level information such as the different filenames of a file and the list of peers providing it.

3 Keyword analysis and paedophile activity.

Keywords play a central role in P2P activity as they are used to search for files (users send keyword-based queries) and to name files. Our dataset contains much information of this kind, with dozens of millions of filenames and queries. Observing the words occurring in these filenames and queries gives information on both the content available in *eDonkey* and the interests of users. The results detailed in [4] are summarized in this section.

First a general statistical analysis of keywords encountered during our measurements was conducted. One interesting observation is that queries contain many specific keywords while filenames are much more generic. It was also observed that there was a huge heterogeneity between keywords: while some appear millions of times, many words appear only a few times. Conversely, the number of keywords entered by each user is very heterogeneous.

Among all observed filenames and queries, approximately 0.1% contains a clear paedophile keyword. This gives an idea of the importance of the phenomenon in *eDonkey*, but one may also notice that other potentially harmful keywords (e.g. *rape* or *torture*) appear much more frequently.

Studying the observed keywords leads to many other interesting results. For instance, the plot in Figure 1 shows that the age of the children involved in paedophile queries or filenames is very low. It also shows that paedophile queries are directed towards content with children significantly younger than the ones claimed in filenames.

Going further, this data was used to derive information on the paedophile nature of keywords. To do so, a set of a few well-known and unambiguous paedophile keywords was selected; then all the filenames or queries containing these words were considered, and the frequency of any word in these filenames and queries was compared to their frequency in other filenames and queries (in particular the ones related to pornographic content). Indeed, the fact that a word appears frequently in a paedophile context was not sufficient to identify paedophile keywords, but if in addition it appears rarely in other contexts then it probably indicate paedophile content. Lists of new paedophile keywords (i.e. different from

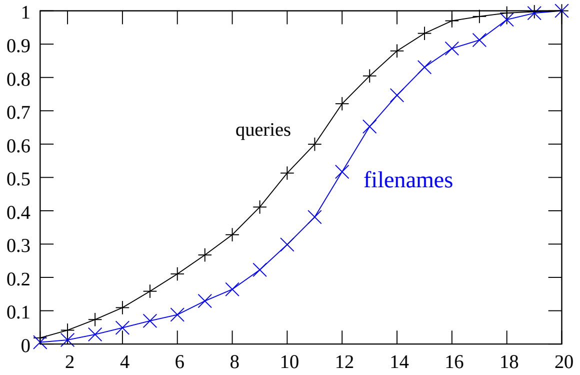

Figure 1: Repartition of ages claimed in filenames and asked for in queries. For each value of n from 1 to 20, all filenames and queries containing the string $n$yo (for n years old) were selected, and for each x the fraction of these strings with $n \leq x$ was plotted. For $n = 10$, about half queries and 40 percent of filenames containing an information about age refer to ages of 10 years old or less. Likewise, $n = 5$ shows that approximately 15% of queries and 7% of filenames containing an information about age refer to ages of 5 years old or less.

the ones we started with) were obtained with this method, which confirms the hypothesis.

Summary of contributions regarding keyword analysis and paedophile activity.
Studying the dozens of millions of keyword-based queries and filenames contained in the captured data shows that the portion of the whole *eDonkey* activity directed towards paedophile content is of the order of 0.1%. Ages indicated in queries and filenames are 5 years old or less in 15% of queries and 7% of filenames. For age 10 the portions are 50% and 40% respectively. The collected data was also used to infer new paedophile keywords based on a set of previously known ones.

4 Content rating and fake detection.

The goal of the Content Rating System (CRS) is to automatically provide a classification of files encountered in *eDonkey* as *porn* or *paedo* files. Several approaches will be followed during the project; at this stage, the use of keywords in filenames only was considered. This provides a ground estimation of what is feasable for future work.

The basic principle is the following. A list of well-identified pornographic and paedophile keywords for which there is no ambiguity is used (*i.e.*, a file having one of these words in its name is considered as a porn or paedo file for sure). For instance, such keywords are *porn* and *childsex*. All the files which have (at least) one of these keywords in their names are then selected, and then the set of all the words encountered in the names of these files is also selected. We define the *porn ratio* (resp. the *paedo* ratio) of each of these words as its number of occurrences in porn filenames (resp. paedo) divided by its number of occurrences in all filenames. This gives a measurement of the porn (resp. paedo) nature of each word. Finally two different ratings of filenames have been defined. The first one is the maximal rating of any word contained in the filename; in this case, a filename which contains at least one word with a strongly porn (resp. paedo) nature will have a high porn (resp. paedo) rating. In the second rating scheme, the average of the ratings of all words appearing in the filename is considered; in this case, a filename with high rating will be a filename in which a large portion of words have a high rating. Both are relevant: the maximum rating scheme takes no risk and considers a worst case situation (a file is pointed out as porn or paedo if there is a hint in this direction); the average rating scheme indicates porn (resp. paedo) files with much more confidence as the filename must be strongly related to this kind of content.

Both maximum and average ratings were included in the Web interface, thus allowing manual inspection and assessment of their relevance. The first tests clearly indicate that they are indeed able to point out pornographic and paedophile files. For more information, see the detailed description of our CRS in [3].

In several cases, however, the efficiency of this CRS is limited by the fact that files have multiple, very different names. Such files are called *fakes*: these files have (at least) a name significantly different from their content. Detecting fakes is a key issue for user protection, as it may help in avoiding unwanted exposition to harmful and/or illegal content. One goal of our project is to develop an automatic Fake Detection System (FDS). Similarly to the CRS, we provide a first version of this FDS based on filenames, which will be improved in further stages of the project.

One may expect to detect fakes simply by counting the number of filenames of each file: a file with many filenames would probably be a fake. Actually, the different names of a file are often simple variations (e.g. changes in separators and case letters, translations, permutations of words, more or less precise descriptions of the content, etc), and so do not indicate a fake. We therefore had to develop a more subtle approach. The key challenge here is to avoid manual inspection and in general manual input (like explicit translation of filenames).

Four statistical indicators were introduced, based on the overlap and difference between the sets of words that compose each filename. For instance, we suppose that two filenames with many words in common are similar, in particular if one of them is included in the others (then, the longest one generally is a more precise description of the content). These four indicators were included in our Web interface for manual investigation and assessment. Although they clearly succeed in providing relevant indication in many cases, they are sometimes inefficient (translations of filenames in particular). Some work is planned to improve this in the project. For more information, see the detailed description of our FDS in [3].

Summary of contributions regarding content rating and fake detection.
Several indicators were proposed to label files as having a pornographic or paedophile nature (content rating system) and/or as having a content significantly different from their names (fake detection system). These indicators are based on keywords encountered in filenames only, but they already give relevant results. They will be improved in future work.

5 Future work.

After one year, the project already obtained significant results. First, very rich data on P2P activity were collected at a scale orders of magnitude larger than previously. This huge volume of data was put in a usable form for researchers and investigators to gain much insight on this activity, in particular paedophile activity. Using this, several keyword analyses brought new information on this activity. A first version of content rating and fake detection system was also set up, which is very promising.

However, much remains to be done to improve current results and reach the project's goals. In particular, we are currently working on refining content rating and fake detection with more subtle techniques; on the time evolution of paedophile keywords to identify new, emerging keywords and tendencies; on larger and longer measurements, focused on paedophile activity; on additional browsing facilities to our Web interface (like keyword-based searches for instance), etc. Finally, a clear view of what occurs on P2P systems, in particular regarding paedophile activity, is expected for next year.

Acknowledgements. This report is a collective work by all participants to the project. It is supported in part by the MAPAP SIP-2006-PP-221003 and ANR MAPE projects.

References

[1] Frederic Aidouni, Matthieu Latapy, and Clémence Magnien. Ten weeks in the life of an edonkey server. Submitted, 2008.

[2] Oussama Allali, Matthieu Latapy, and Clémence Magnien. Measurement of edonkey activity with honeypots. Submitted, 2008.

[3] Matthieu Latapy, Clémence Magnien, and Guillaume Valadon. First report on database specification and access including content rating and fake detection system. `http://antipaedo.lip6.fr/`.

[4] Clémence Magnien, Matthieu Latapy, Jean-Loup Guillaume, and Bénédicte Le Grand. First report on paedophile keywords observed in edonkey. `http://antipaedo.lip6.fr/`.

[5] Collective work. Measurement and analysis of p2p activity against paedophile content – presentation of the project. `http://antipaedo.lip6.fr/`.

[6] Collective work. Report on current knowledge regarding paedophile activity in p2p systems. `http://antipaedo.lip6.fr/`.