Measurement and Analysis of P2P Activity Against Paedophile Content

General Public Report

Fall 2009

http://antipaedo.lip6.fr

Coordinator: **Matthieu Latapy**

Matthieu.Latapy@lip6.fr

Abstract

Peer-to-peer (P2P) systems are nowadays widely used to exchange files, and it is acknowledged that they host much paedophile activity. However, current knowledge of this specific activity remains very limited, and almost no tool exist for user protection. Likewise, tools and knowledge for policy making and law enforcement are far from sufficient. The goal of the *Measurement and Analysis of P2P Activity Against Paedophile Content* project is to improve this situation significantly by conducting large-scale measurements of P2P activity, with a focus on paedophile activity. Obtained data is then used for analysis of this activity, thus leading to key insights. The project also designs and implements tools for user protection and law enforcement institutions, in particular automatic detection tools for paedophile keywords and queries, as well as content rating and fake detection systems. The goal of this report is to give a synthetic general-public view of the obtained results.

1 Introduction

Many studies and independent contributions show that a huge amount of paedophile [1] and more generally harmful contents are distributed using peer-to-peer (P2P) file exchange systems. This can easily be checked by anyone, since a simple query on the keywords *porn* or *pedo* with a classical P2P client leads to hundreds, and up to several thousands, of answers.

The presence of such contents and their very easy access is worrying for P2P users, in particular children. Indeed, a significant number of children, in particular teenagers, nowadays use P2P systems and may therefore be exposed to such contents.

Despite the fact that this situation is widely acknowledged, there is still no satisfactory technique to protect P2P users, in particular children, from paedophile and other harmful contents. Similarly, only few tools exist to help law enforcement authorities and other child protection organisations in fighting P2P paedophile exchanges. Actually, there is

[1] In this report, the term *paedophilia* refers to child sexual abuse in general, including hebephilia, a child being defined as any individual aged less than 18 years. This terminology mixes together different aspects of the problem, but our methodology does not in general allow to distinguish between them. Figure 1 below gives some insight on this, though.

even an important lack of precise knowledge on this topic. It has been observed at many occasions that this has a deep impact on our ability to fight these exchanges.

The objective of the *Measurement and Analysis of P2P Activity Against Paedophile Content* project is to tackle these issues by designing and implementing key software, setting up reference databases and conducting leading studies. Our aims are both to help in protecting P2P users, in particular children, and help law enforcement authorities and other child protection organisations in their tasks.

This report presents our actions and findings to the general public. We focused on four tasks: large-scale and long-term measurement of activity in a large P2P system (*eDonkey*) with a focus on paedophile activity; design and implementation of keyword and query detection tools useful for filtering, data inspection and law enforcement; design and implementation of a content rating and fake detection tool for user protection and for helping manual classification by law enforcement personnel; and analysis of the obtained data with the tools we designed in order to improve significantly current knowledge of paedophile activity.

The project consortium gathers a multi-disciplinary set of European research institutions and NGO, funded by the European Commission, national agencies, and NGOs (see acknowledgements at the end of this report). The project is conducted in close collaboration with law enforcement institutions, which helps defining relevant priorities and assessing obtained results. See our website (`http://antipaedo.lip6.fr`) for details.

2 Measurement and data

Because of their huge size (typically millions of users and files), their distributed nature (no central server or administration), their high dynamics (peers and files join and leave the system over time), and various other difficulties (like poor protocol documentation and policy constraints, for instance), measurement of P2P systems is a challenging task. Data collected before the project were very limited in size, in duration of the measurement and in quality. We therefore had to design and conduct measurement strategies able to provide data which fulfils our needs.

In order to avoid dispersal of efforts, the project focused on one specific P2P system, called *eDonkey* (it is also sometimes called *eDonkey2000*, *eD2k*, or *Emule*). This system has several advantages which justify our focus. First, it is among the most widely used systems currently in use, with millions of users and millions of available files, and this has been true for several years. In particular, it is a general-public system, with many non-expert users. We therefore expect to observe a significant portion of P2P exchanges in it. Its protocol, although poorly documented, is now quite well known, which is also important for the design of measurement tools. Finally, it has a semi-distributed structure: it relies on a set of servers recording which files are available and who provides them (but not file contents), and peers which host and exchange files. This makes it possible to conduct server-side measurements (see below), which plays a key role in our context.

Measurement methods

Although the *eDonkey* protocol has many variants and subtle features, 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, filenames, 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 the chosen files. Afterwards, the peer software directly contacts the providers to get the files.

Several complementary approaches are possible to observe activity in *eDonkey*. Each has its own advantages and drawbacks, therefore we used several. We outline below the different techniques we used, and summarise the main features of the obtained datasets. Full details are available in cited publications.

Server-side measurements [1]. A capture program may be placed on a server to record the queries it receives and the answers it sends. This kind of measurement therefore is similar to the measurements conducted by Web search engines like *Google*, which record the queries sent by users and the answers they obtained. This way, *all* queries managed by this server may be captured, but we cannot know whether users actually exchange files or not.

Measurement by client sending queries [8]. A client program sends queries to servers using a set of predefined keywords to monitor, and records the obtained answers (lists of files and providers for these files). 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 (i.e. the websites which provides pages containing the chosen keywords). This may be repeated periodically during long periods to obtain more data.

Measurement by honeypots [2]. A client program may advertise some files of interest (by declaring to servers that it owns these files) and then record the queries it receives from other peers who seek these files. 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.

Notice that these different measurement approaches provide 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; honeypots capture which clients actually attempt to download files, which cannot be observed by the other methods.

Obtained data

Let us first insist on the fact that, for privacy protection concerns, all the data we collected 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 contents: only *queries* and the corresponding answers have been recorded. This information is very rich, as it captures both user behaviours and the kind of files exchanged. It would not be possible to collect information at this scale by observing file contents.

We conducted two **server-side measurements** [1]. The first one has been made

in 2007 on a large server (at that time) during a period of 10 weeks. This led to the observation of approximately 89 million peers and 275 million files. We recorded *all* exchanges between the server and these peers; these exchanges can be separated into the four types, corresponding to the four steps of the communication scheme presented above. This represents 117 million keyword-based queries (type 1) and 24 millions filenames (type 2), as well as information on which peers provide or seek which files (types 3 and 4).

We performed our second server-side measurement in 2009 on a medium-sized server during a period of 15 weeks. This time, only recording of keyword-based queries was possible, but geolocation of users was performed. We observed 106 million queries from 16 million users.

We conducted a **measurement by client sending queries** which lasted for 7 months [8]. During this period of time, the client sent every 12 hours 8 paedophile queries and 7 non-paedophile queries to approximately 100 *eDonkey* servers. We observed this way 3 million files (among which almost 800 000 were paedophile) and 3.5 million users (among which 1.3 millions provided at least one paedophile file). As expected, because this measurement is focused on paedophile activity, we observe a lot of such activity.

We conducted several **honeypot measurements** [2], including a one-month long one in which the clients advertised 32 files and observed 24 649 peers. We also conducted a measurement aimed at collecting as much data as possible, based on a unique honeypot; it led to the observation of more than 250 000 files and 850 000 peers. However, as honeypot measurements interfere with law enforcement activities, we did not use them further.

Finally, the obtained datasets are orders of magnitude larger than previously available ones. One significant contribution of the project consists in the public provision of these fully anonymised datasets to the research community, and datasets containing more information for partners.

Data access

Gaining insight on activity of millions of users exchanging hundreds of millions of files is extremely difficult. In order to help for this task we developed a **Web interface** to browse our data and get more intuition regarding its content [12]. We first computed key information like the number of files each peer provided, the time at which it first appeared, the queries it entered, the names of each file, etc. We then included it in a system which generates a web page for each observed file and each peer, with all available information on it. In addition, it is possible to browse the data from files to related peers, and conversely. Further information (like content rating and fake detection facilities, see Section 4) is also available from this interface, as well as a facility for finding directly paedophile queries (see below, Section 3). This gives a convenient way to explore the dataset and develop an intuition on its content. See [12] for full details.

Summary of results 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 it. This interface also provides higher level information such as the different filenames of a file and the list of peers providing it.

3 Detection of paedophile keywords and queries

Keywords play a fundamental role in P2P activity since they are used to send queries to the system and to describe file content in filenames. As a consequence, knowledge of paedophile keywords is a crucial matter for monitoring exchanges of paedophile contents, for filtering purposes and for data inspection. For these reasons, an important part of our activity was dedicated to the study of paedophile keywords.

Some people interested in paedophile content use confidential keywords to avoid detection. These keywords do not seem to be paedophile to a casual observer (they often even seem meaningless). As such keywords tend to become more widely known over time (by law enforcement personnel in particular), new keywords may appear. One may also guess that paedophile users who forge such keywords try to make them difficult to detect, and change them frequently to keep them secret.

Ensuring accurate and up-to-date knowledge of these keywords is therefore challenging. This knowledge nowadays relies on manual inspection by experts, which is highly time consuming. In order to improve this situation we explored the possibility of designing automatic detection tools to help in this task [3]. We compared several methods (including some designed within the project [3, 4, 13]), and submitted obtained results to experts. Our conclusion is that **automatic methods may help greatly in maintaining accurate lists of paedophile keywords**. In addition, we produced this way lists of very relevant paedophile keywords, which is a significant contribution in itself (some keywords were previously unknown).

We then used these lists and manual inspection of huge datasets to design a tool for **automatic detection of paedophile queries** [10]. This tool tags a query as paedophile if it contains a specific paedophile keyword or a combination of other keywords (mainly keywords related to sexuality and keywords related to childhood or age indications).

An in-depth assessment of the performances of this tool by experts of the field led to the conclusion that it is wrong in only approximately 1% of cases when it tags a query as paedophile, and misses only approximately 25% of all paedophile queries. As the fraction of paedophile queries is very low, these results are excellent. Using it, we obtained for the first time **a list of hundreds of thousands of paedophile queries** found in the data collected in the project. This list is helpful in improving knowledge of paedophile activity. See [10] for details.

Summary of contributions regarding paedophile keywords and queries.
We produced two key tools regarding paedophile keyword analysis: automatic detection of paedophile keywords, and automatic detection of paedophile queries. Both are significant improvements of the state of the art. They play a crucial role in most studies performed in the project and are helpful in many tasks raised by fighting against paedophile activity. In addition, they led to the construction of lists which have their own interest: a list of specific paedophile keywords and a list of paedophile queries, both usable by law enforcement personnel for investigation, by search engines for filtering, and by researchers for mining large datasets in search for paedophile activity.

4 Content rating and fake detection

Because many files in P2P systems have pornographic or paedophile content, and because of the presence of many *fakes* (files with content significantly different from their name), users (including children) may be unintentionally exposed to such contents. In addition to the shock experienced by most users in this situation, it is suspected that unwanted access to paedophile content may increase interest in it.

Moreover, law enforcement institutions maintain large databases of files known as being paedophile, which is highly resource consuming. Likewise, Internet service providers (as well as administrators of *eDonkey* servers) implement filtering techniques which rely on such databases.

In this context, a system able to **automatically point out files which probably have a paedophile or pornographic content** would be of high interest.

In many cases, in particular in P2P systems, filenames are not available or may be misleading, and the number of files to handle is too large to download them and inspect their content. We therefore explored a new approach [7] consisting in using the information we have on user interests to infer file types: if a given user provides two different files, then these files are somewhat related. They are even more related if many users provide them both. Using the large amount of data collected within the project, we follow this approach to build a structure among files which captures their similarities according to user interests. We then used a clustering method developed in the project to construct groups of similar files. In particular, this procedure groups together files with paedophile nature.

We obtain this way an automatic rating of each file as probably paedophile, maybe paedophile, or not paedophile. Similar results hold for pornographic contents. The relevance of these ratings has been demonstrated by observing the classification of files known to have paedophile and pornographic content. See [7] for details.

Using these ratings and filenames, we also provide a **fake detection system**: a file is classified as fake if its rating indicates paedophile or pornographic content but its name does not, or conversely. Again, confrontation of our results to known data showed the relevance of this classification [7].

The obtained content rating and fake detection systems were incorporated into our Web **interface** to browse the data [12], described above, in order to both help in inspect-

ing the data and make it easier to assess the relevance of its results. Importantly, we also implemented an interface which makes it easy to incorporate this system to P2P client software, in order to warn users trying to download suspicious content.

Summary of contributions regarding content rating and fake detection.
We designed and implemented an original approach able to classify a file as probably paedophile, maybe paedophile, or not paedophile, without inspecting its content nor its name; instead, we use the fact that files are provided by several users to infer their nature. This tool is useful for law enforcement and filtering, which use lists of files known to be paedophile. We obtained similar results for pornographic files. We then used these ratings for fake detection, thus providing an important tool for user protection. A web interface to this tool makes it easy to include it in client software.

5 Improved knowledge of paedophile activity

Despite the fact that paedophile activity in P2P systems is widely acknowledged as alarming, knowledge about this activity remains very limited. However, such knowledge is crucial for understanding what is precisely going on, help law enforcement in their fight against this type of activity, and for policy making (including resource allocation and Internet regulation).

One key goal of the project was to improve this situation by **deriving accurate and reliable information on paedophile activity** in the *eDonkey* system, which we consider as representative. We focused in particular on the rigourous evaluation of the fraction of paedophile queries among all queries managed by the system, the fraction of users involved in such queries [9], as well as the identification of various kinds of files and users, and several other basic statistics [6]. We also observed the time evolution of these quantities and the structure of paedophile activity (including its geographical distribution) [5, 11].

Quantification and evolution

Our first goal was to obtain, for the first time, an accurate and rigourous quantitative estimate of the *amount* of paedophile activity in a large P2P system [9]. We focused on the fraction of paedophile queries entered in the system, and the fraction of users who entered these queries. We first computed the fraction of queries tagged as paedophile by our automatic paedophile query detection tool (see above and [10]). As the rates of error are known for this tool, we then inferred our estimate: **approximately two out of one thousand observed queries are paedophile queries**.

Estimating the fraction of *users* entering paedophile queries, which we call paedophile users, is much more difficult. Indeed, several users may use the same Internet address, and conversely a given user may use several addresses. Using only addresses would lead to a significant overestimate of the fraction of paedophile users, which we demonstrated. Using another technical information (the connection port), we however observed that we

were able to distinguish distinct users in most cases, leading to an underestimate of the fraction of paedophile users. We finally conclude that **approximately two users out of one thousand enter paedophile queries**. See [9] for details.

These estimates were computed on datasets collected in 2007 and 2009, and we obtained similar values. Moreover, we studied their evolution during periods of several months; no significant variation was observed at this level, thus indicating that **proportions of paedophile activity are very stable**. Likewise, the use of each specific keyword remained very stable all along the measurements.

Instead, a daily variation was observed, with a decrease of paedophile activity between 5pm and 11pm [5, 9]. This may indicate that, unlike often thought, users seeking paedophile content are involved in casual social life (work, family, etc). Investigating this, by comparing it to queries for pornographic content for instance, is an interesting perspective.

User interests

Our work on quantification of paedophile activity showed that thousands of users enter paedophile queries; our automatic paedophile query tool makes it possible to locate these queries in the huge amount of data collected and to inspect them further. They provide much information on user interests and behaviours, which is crucial to improve our understanding of what is going on regarding paedophile activity in P2P systems [6].

First, many queries contain **age indications**, for instance of the form *n yo* for a given number *n*, the suffix *yo* standing for *years old*. Filenames also often contain such age indications. Focusing on paedophile queries and filenames (as detected by our automatic tool), we inspected the age distributions in queries and filenames, see Figure 1 and [6].

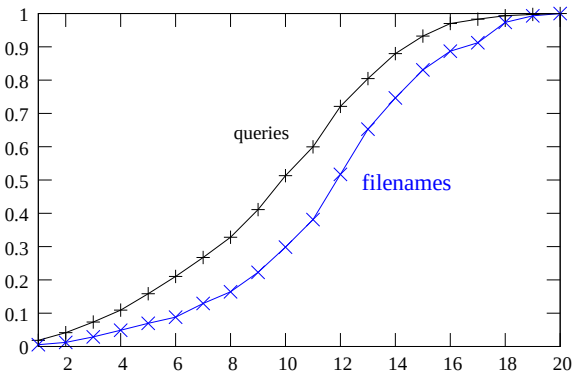

Figure 1: Repartition of ages claimed in paedophile filenames and asked for in paedophile queries. For each value of n from 1 to 20, all such filenames and queries containing the string *n yo* were selected, and for each x on the horizontal axis we plot the fraction containing an age $n \leq x$. For $x = 10$, we obtain that about half queries and 40 percent of filenames containing an age information refer to ages of 10 years old or less. Likewise, $n = 5$ shows that approximately 15% of queries and 7% of filenames refer to ages of 5 years old or less.

These analyses show clearly that **a significant proportion of queries and filenames refer to very young children**, see Figure 1. In addition, queries generally

indicate younger ages than filenames; this means that **there is a demand for younger ages than what is actually offered**.

Going further, we studied for each relevant user the proportion of paedophile queries he/she entered. This highlighted the fact that **some users focus on paedophilia** in their use of the system. We observed wide ranges of different behaviours, though: from users who entered a single paedophile query but a large number of non paedophile queries, to users who entered exclusively, and many, paedophile queries. See [6] for details.

Finally, we inspected in more details series of queries entered by some specific users [6], thus showing that different profiles exist and that our data and tools succeed in providing means of examining them. This is one of the main perspectives of our work, as explained below.

Maps

Policy-making and law enforcement institutions generally operate at the national level, or at least at a regional level. Therefore, **geolocated analysis of paedophile activity** is of prime interest. Using geolocated recordings available in our second server measurement, we computed the number of queries and the fraction of paedophile queries for each country (when available data was sufficient to make such statistics significant) [11].

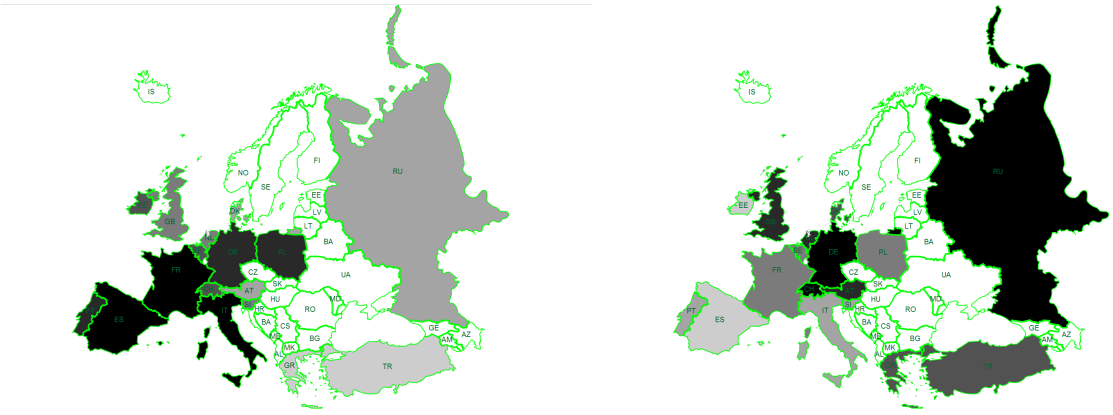

Figure 2: Maps of captured activity and paedophile activity in Europe. Left: colours represent the number of queries observed; right: they represent the fraction of paedophile queries (for countries for which this statistics is relevant; others are in white). Dark colours indicate high numbers, light ones low numbers. It appears clearly that south-west countries submitted many of the queries we captures, but that the fraction of paedophile queries in these sets is higher for eastern and central countries (except Poland) and England.

This led us to the conclusion that **the situation varies strongly between countries**, with fractions of paedophile queries orders of magnitude larger in some countries than in others. In particular, the observed fraction of paedophile queries from Russia is twice larger than the fraction from the second most prolific country (USA). In addition, the countries which use *eDonkey* the most are not necessarily the ones with highest interest in paedophile content, see Figure 2 and [11]. This may indicate countries where more attention should be paid to this issue.

Summary of contributions regarding improved knowledge of paedophile activity.
Thanks to data and tools from the project, we were able to derive, for the first time at this level of accuracy and reliability, quantitative estimates of the fractions of paedophile queries in the system and users responsible for them: both are close to 2‰. We then examined user interest in more details. Age indications in paedophile queries and filenames reveal that a significant proportion of them refer to very young children, and that there is a demand for younger ages than what is actually offered. Going further, we observed that some users strongly focus on paedophile content. We also examined the geographical distribution of paedophile activity in *eDonkey*, showing that the situation varies between countries.

6 Conclusion and perspectives

The project collected huge amounts of data, orders of magnitude larger than the ones previously available. We also designed and implemented key tools for inspecting paedophile activity in such data: an automatic paedophile keyword detection tool and an automatic paedophile query detection tool. These tools are useful in themselves, and they made it possible to produce data on paedophile activity at an unprecedented level of size, precision, and reliability. Using them, we also constructed a content rating and fake detection system useful for law enforcement and user protection. Finally, we conducted in-depth and rigourous analyses which provide much insight on paedophile activity. In particular, the proportion of paedophile queries entered in *eDonkey* is close to 2 ‰, as well as the proportion of users who entered them.

These contributions may be deepened in several directions. First, similar measurements and analyses may be conducted on other P2P systems, or even on other kinds of data exchange media. Comparing paedophile activity in different systems would probably shed more light on this phenomenon. Clearly, the methods and tools developed in this project would be of great help in inspecting paedophile activity in these other contexts.

Likewise, much remains to be done regarding the analysis of user behaviours, and the project provides data and tools for this. First, the topics of interest of paedophile users may be explored, as well as the way their interest in paedophile content evolves over time. Indeed, a key question is whether this interest evolves towards younger ages or not, and if some topics tend to conduct people to develop an interest in paedophile content. Inspecting long series of queries entered by paedophile users may help in answering these questions, and in identifying profiles of particular interests.

Finally, this project demonstrated the feasibility and interest of large-scale measurements and analysis of paedophile activity. However, it is only a first step towards a better understanding and monitoring of such activity and much remains to be done.

Acknowledgements. This report is a collective work by all participants to the project. We warmly thank the administrator of the `peerates.net` *eDonkey* server for his help in collecting data. We also thank the experts who helped in assessing the results, in particular Philippe Jarlov who also contributed significantly to data collection. This

work is supported in part by the MAPAP SIP-2006-PP-221003 project, ANR MAPE project, and *Action Innocence Monaco* NGO.

All the project technical reports cited in this text are available online at the project website.
`http://antipaedo.lip6.fr`

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] Oussama Allali, Matthieu Latapy, and Clémence Magnien. Measurements of *eDonkey* activity with distributed honeypots. In *Proceedings of HotP2P'09*, 2009.

[3] Christian Belbèze, David Chavalarias, Ludovic Denoyer, Raphaël Fournier, Jean-Loup Guillaume, Matthieu Latapy, Clémence Magnien, Guillaume Valadon, Vasja Vehovar, and Aleš Žiberna. Technical report on *Automatic Identification of Paedophile Keywords*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

[4] 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.

[5] Bénédicte Le Grand, Jean-Loup Guillaume, Matthieu Latapy, and Clémence Magnien. Technical report on *Dynamics of Paedophile Keywords in* eDonkey *queries*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

[6] Jean-Loup Guillaume, Matthieu Latapy, Bénédicte Le Grand, and Clémence Magnien. Technical report on *Behaviours of Users Entering Paedophile Queries*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

[7] Jean-Loup Guillaume, Matthieu Latapy, Clémence Magnien, and Guillaume Valadon. Technical report on the *Content Rating and Fake Detection System*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

[8] Philippe Jarlov, Matthieu Latapy, Frédéric Aidouni, Clémence Magnien, Christophe Berger, and Firas Bessadok. Monitoring paedophile activity in a P2P network. *Forensic Science International*, 2009.

[9] 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.

[10] 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.

[11] Matthieu Latapy, Clémence Magnien, Raphaël Fournier, and Massoud Seifi. Technical report on *Maps of Paedophile Activity*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

[12] 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.

[13] Vasja Vehovar, Aleš Žiberna, Matej Kovačič, Andrej Mrvar, and May Doušak. Technical report on *An Empirical Investigation of Paedophile Keywords in P2P Activities*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

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