Comparing Pedophile Activity in Different P2P Systems
Raphaël Fournier, Thibault Cholez, Matthieu Latapy, Isabelle Chrisment, Clémence Magnien, Olivier Festor, Ivan Daniloff

To cite this version:

HAL Id: hal-01052773
https://inria.hal.science/hal-01052773
Submitted on 28 Jul 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Comparing Pedophile Activity in Different P2P Systems

Raphaël Fournier 1,*, Thibault Cholez 2, Matthieu Latapy 3,4, Isabelle Chrisment 2, Clémence Magnien 3,4, Olivier Festor 2 and Ivan Daniloff 3,4

1 L2TI/Institut Galilée, Université Paris-Nord, 99 avenue JB Clément, 93430 Villetaneuse, France
2 LORIA/INRIA Nancy-Grand Est, 615 Rue du Jardin Botanique, 54600 Villers-lès-Nancy, France;
   E-Mails: Thibault.Cholez@inria.fr (T.C.); Isabelle.Chrisment@loria.fr (I.C.);
   Olivier.Festor@loria.fr (O.F.)
3 Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005 Paris, France;
   E-Mails: Matthieu.Latapy@lip6.fr (M.L.); clemence.magnien@lip6.fr (C.M.);
   ivan.daniloff@laposte.net (I.D.)
4 CNRS, UMR 7606, LIP6, F-75005 Paris, France

* Author to whom correspondence should be addressed; E-Mail: raphael.fournier@univ-paris13.fr;
   Tel.: +33-149-402-824.

Received: 19 March 2014; in revised form: 15 May 2014 / Accepted: 16 May 2014 / Published: 1 July 2014

Abstract: Peer-to-peer (P2P) systems are widely used to exchange content over the Internet. Knowledge of pedophile activity in such networks remains limited, despite having important social consequences. Moreover, though there are different P2P systems in use, previous academic works on this topic focused on one system at a time and their results are not directly comparable. We design a methodology for comparing KAD and eDonkey, two P2P systems among the most prominent ones and with different anonymity levels. We monitor two eDonkey servers and the KAD network during several days and record hundreds of thousands of keyword-based queries. We detect pedophile-related queries with a previously validated tool and we propose, for the first time, a large-scale comparison of pedophile activity in two different P2P systems. We conclude that there are significantly fewer pedophile queries in KAD than in eDonkey (approximately 0.09% vs. 0.25%).

Keywords: P2P networks; eDonkey; pedophile activity
1. Introduction

Pedophile activity is a crucial social issue and is often claimed to be prevalent in peer-to-peer (P2P) file-sharing systems [1,2]. However, current knowledge of pedophile activity in these networks remains limited.

Recently, research works have been conducted to improve this situation by quantifying pedophile activity in Gnutella and eDonkey, two of the main P2P systems currently deployed [3,4]. They conclude, respectively, that 1.6% and 0.25% of queries are of pedophile nature, but these numbers are not directly comparable as the authors use very different definitions and methods. Such comparisons are of high interest though, since differences in features of P2P systems, such as the level of anonymity they provide, may influence their appeal for pedophile users.

In this paper, we perform a comparison for the first time. We focus on the KAD and eDonkey P2P systems, which are the names given to the two underlying P2P networks used by the popular eMule file-sharing application. They are both widely used, accounting together for almost 10% of the global Internet traffic in Europe in 2012 [5], but they differ significantly in their architecture: while eDonkey relies on a few servers, KAD is fully distributed. This lack of centralization may lead users to assume that KAD provides a much higher level of anonymity than eDonkey. Comparing the two systems sheds light on the influence of a distributed architecture on pedophile behavior and increases general knowledge on pedophile activity in P2P systems.

The term pedophilia is popularly used to denote adult sexual engagement with children, both prepubescent and pubescent. The definition of pedophilia we use in this article thus encompasses both the medical definition of pedophilia (sexual interest in prepubescent children) and hebephilia (sexual interest in pubescent children not sexually mature).

We discuss related work in Section 2, to give an overview of the state-of-the-art on online pedophile activity detection and analysis. Section 3 presents a short introduction to P2P systems, before our description of our datasets and how we collected them (Section 4). We then present the details of our comparison of the amount of pedophile queries in KAD and eDonkey in Section 5. Section 6 focuses on an important feature of pedophile activity: ages entered in queries. Finally, in Section 7, we introduce a methodology to estimate the fraction of pedophile queries in KAD from the one in eDonkey.

2. Related Work

Collecting P2P traces is an active topic for years, but it is mostly aimed at analyzing peer behavior to help with future P2P protocol design. In 2006, authors of [6] and [7] explored the social and technical issues related to online child pornography and opened the way to the research in the field. The first detailed quantitative study focusing on a P2P system was proposed in [3], using an active methodology (sending specific queries and analyzing the answers provided by the search engine). Since then, several approaches have been proposed to gage the extent of the phenomenon. Among them, [8] presented filename categorization tool, while [9,10] proposed to label suspicious chat conversations. [11] especially analyzed aged-related queries.

A first large-scale study of P2P search-engine queries was presented in [12]. Their study focused on “onset”, the first deliberate viewing of child pornography. They gathered the Top 300 queries submitted
to the popular Isohunt tracker (part of the BitTorrent network) and published on the website Isohunt.com. Their study lasts for 3 months, a scale similar to ours, but they resort to manual classification of the queries. Their dataset is particular, as it only gives a relative popularity order for some queries, and may not provide any indication on the extent of child pornography in the network. Plus, with only 300 queries collected daily, they get very few pedophile queries (only 3), which leads to results with a limited statistical significance. However, their discussion is truly interesting, including comments on whether “regular pedophile users” are likely to submit several times the same query (to “build a collection”), while first-time users may not (they do not progress to downloading material once they have discovered the meaning of intriguing pedophile sequences such as *pthc*). This bias may lead query-based studies like ours to slightly overestimating the demand for child pornography, and would impact estimations on the number of pedophile users, but additional filtering based on the IP-address or the client ID could limit this issue.

In [4], the authors developed and assessed a dedicated tool for search engine query classification, and collected large-scale datasets on eDonkey (up to 28 weeks of uninterrupted experiment). We use here their tool and one of their datasets. Part of their work was later reused by another team to study another P2P network, BitTorrent [13]. The European Commission has set up a “Safer Internet” program [1], which funded some large research projects such as MAPAP [14] and iCOP [15].

In parallel, authors of [16,17] provided an extensive study (one-year long) on child pornography on Gnutella and eMule, partnering with law enforcement to develop software platforms and collect data on child pornography trafficking. They made a precious contribution to understand the “supply”: how many users are involved in the distribution of files, what are their importance in the network, etc. In [16], they evaluated different strategies to best fight pedophile activity given the limited resources of law enforcement and proposed an efficient metric to target the most prominent peers.

While having a smaller scale, our study is the first to provide a methodology to gain new knowledge from the proper comparison of data collected from two P2P networks which architecture and monitoring capacity are totally different. Moreover, if the general user behavior in the KAD network was detailed in [18], our article is the first to study whether its decentralized architecture is prone to favor criminal activity.

### 3. P2P Systems

P2P systems are computer networks in which every user may share content with others members. They have become popular because they gather large amount of digital contents (books, movies, music) which can be obtained for free. Copyrighted material is available (however not authorized) and pornography is widespread. Accessing a P2P network is generally easy: a user only needs to download and install on his computer a single application, which will handle the connection process to the network. Then, he can search for files with some keywords, and gets a list of corresponding available files. The application sends messages to the network to find providers of the selected files, and then users interconnect directly to exchange them.
P2P networks are easy to access for both providers and consumers. Contents are obtained free of charge, and rather anonymously (no personal details are required). These features make such networks appealing for illegal activities such as pedophile material trafficking.

P2P networks account for approximately a fifth of the global bandwidth use on the Internet. Bittorrent is the most prominent P2P network nowadays, preceding eDonkey and KAD (the usage of which decline in Europe). For instance, an important eDonkey server received on average 8.8 million queries per week between 2009 and 2012 [4].

4. Experimental Setup and Datasets

In order to compare pedophile activity in two different P2P systems, we first need appropriate datasets, the collection of which is a challenge in itself. In KAD and eDonkey, different kinds of measurements are possible, depending on the details of the network’s architecture.

In eDonkey, servers index files and providers for these files, and users submit keyword-based queries to servers to seek files of interest to them [19]. By monitoring such a server, one may collect all those queries [20]. Here, we record all queries received by two of the largest eDonkey servers during a three-month period in 2010. The servers are located in different countries (France and Ukraine) and have different filtering policies: the French server indexes only non-copyrighted material, while the Ukrainian server openly indexes all submitted files. Monitoring two such different servers will allow us to compare them in order to know if server policy impacts our results.

To collect KAD data, we use the HAMACK monitoring architecture [21], which makes it possible to record the queries related to a given keyword by inserting distributed probes close to the keyword ID onto the KAD distributed hash table. We supervise 72 keywords, which we choose to span well the variety of search requests entered in the system, with a focus on pedophile activity: a set of 19 paedophile keywords (babyj, babyshivid, childlover, childporn, hussyfan, kidzilla, kingpass, mafiasex, pedo, pedofilia, pedofilo, pedoland, pedophile, pthc, ptsc, qaqaazz, raygold, yamad, youngvideomodels), which are known to be directly and unambiguously related to pedophile activity in P2P networks; a set of 23 mixed keywords (1yo, 2yo, 3yo, 4yo, 5yo, 6yo, 7yo, 8yo, 9yo, 10yo, 11yo, 12yo, 13yo, 14yo, 15yo, 16yo, boy, girl, mom, preteen, rape, sex, webcam) frequently used in pedophile queries but also in other contexts (for instance, Nyo stands for N years old and is used by both pedophile users and parents seeking games for children of this age); and a set of 30 not paedophile keywords (avi, black, christina, christmas, day, doing, dvdrrip, early, flowers, grosse, hot, house, housewives, live, love, madonna, man, new, nokia, pokemon, rar, remix, rock, saison, smallville, soundtrack, virtual, vista, windows, world) used as a test group and a priori rarely used in pedophile queries. The sets of keywords were established using the work on pedophile query detection presented in [4]. Notice that our set of keywords contains mainly common English words (love, early, flowers), but some are in other languages (saison, pedofilia), and some are also brand names (pokemon, nokia).

Because of the differences in architectures of the two networks and of the measurement methodologies, we obtained very different datasets, which are not directly comparable: in eDonkey, we observe all queries from a subset of users whereas in KAD we only observe queries related to a given keyword, but from all users. In addition, based on various versions of KAD clients, the measurement
tool only records the queries containing a monitored keyword placed in first position or being the longest in the query. As a consequence, with a short keyword such as avi, a name extension for video files, we almost only record queries in which it is the unique keyword, because otherwise it most likely is neither the longest nor the first word in any query. In order to obtain comparable datasets, we therefore limit our study to a subset of our datasets: the queries composed of exactly one word among the 72 keywords we monitor.

As a result of this construction process, we obtain three datasets, which we call eDonkeyFR, eDonkeyUA and KAD. They contain 241,152, 166,154 and 250,000 queries respectively, all consisting of a unique keyword from our list of 72 monitored keywords, which ensures that they are comparable. The server corresponding to the eDonkeyFR dataset is located in France, while the one corresponding to eDonkeyUA is in Ukraine. Their large sizes make us confident in the reliability of our statistical results presented hereafter.

5. Amount of Pedophile Queries in eDonkey versus KAD

The most straightforward way to compare the pedophile activity in different systems certainly is to compare the fraction of pedophile queries in each system. Figure 1 presents the fraction of queries for each category of keywords. This plot clearly shows that there are very distinct search behaviors in the two networks, since values obtained for the paedophile and not paedophile categories significantly differ between KAD and the two eDonkey datasets. More surprisingly, the fraction of pedophile queries is significantly lower in KAD than in eDonkey which is in sharp contradiction with previous intuition, as KAD is assumed to provide a higher level of anonymity. The plot also shows that values obtained for the two eDonkey servers are similar, which indicates that very different filtering policies have no significant influence on the amount of pedophile queries.

In order to gain a more detailed insight on this phenomenon, we study the frequencies of each keyword separately in the three datasets. As we want to explore possible correlations between the pedophile nature of a keyword and its frequency, we need a way to quantify the pedophile nature of a keyword. To do so, we use the 28-week dataset and the pedophile query detection tool from [4], which divides a dataset between paedophile and not paedophile queries (with a precision above 98% and a recall above 75%). We denote by Q the whole dataset of queries, and by Q(k) the set of queries containing a given keyword k. For each keyword k, we obtain Q(k) = N(k) + P(k), where N(k) and P(k) are the subset of queries containing keyword k and tagged as not paedophile or paedophile, respectively. We then define the paedophile coefficient \( \pi(k) \) of keyword k as: \( \pi(k) = \frac{|P(k)|}{|Q(k)|} \). If all the queries with keyword k are pedophile queries, \( \pi(k) = 1 \), and if none of them are, \( \pi(k) = 0 \). All keywords in the not paedophile category have a paedophile coefficient below 0.006. For keywords in the mixed category, the paedophile coefficient is above 0.01 and below 0.4. All paedophile keywords have a paedophile coefficient above 0.885. Finally, we plot in Figure 2 the ratios \( \frac{f_{eDonkeyFR}(k)}{f_{kad}(k)} \) and \( \frac{f_{eDonkeyUA}(k)}{f_{kad}(k)} \), where \( f_s(k) \) denotes the frequency of queries composed of keyword k in the dataset s, for each of our 72 keywords. We rank keywords on the horizontal axis in increasing order of paedophile coefficient. The horizontal line represents \( y = 1 \), which enables a visual comparison of the values: if the point is below the line, then the keyword is more frequent in KAD, otherwise it is more frequent in the eDonkey dataset.
**Figure 1.** Fraction of queries of each kind in our three datasets.

![Graph showing fraction of queries for not paedophile, mixed, and paedophile categories across datasets KAD, eDonkeyFR, and eDonkeyUA.](image)

**Figure 2.** Ratio of keyword frequencies in eDonkey vs. KAD. Keywords are ranked in increasing order of paedophile coefficient. Points above the $y = 1$ horizontal line indicate keywords more frequent in the corresponding eDonkey dataset; below the line keywords are more frequent in KAD.

![Graph showing ratio of keyword frequencies across eDonkeyFR vs KAD and eDonkeyUA vs KAD.](image)

This plot gives a clear evidence for a correlation between the pedophile nature of a keyword and its higher presence in eDonkey than in KAD. In addition, the frequencies in both eDonkey datasets are very similar for the vast majority of keywords.
We therefore conclude that anonymity is not the prevailing factor when pedophile users choose a network, since neither the decentralized architecture of KAD nor the different filtering policies increase the frequency of pedophile queries. Instead, the frequency of pedophile queries is even higher in eDonkey than in KAD. Finding an explanation for this unexpected phenomenon is still an open question. The higher technical skills required to use KAD may be part of the explanation. Users may also search content on eDonkey while protecting their privacy with other tools, such as Virtual Private Networks or TOR [22]. The fact that in KAD search requests are sent over UDP and cannot benefit from TOR anonymization could explain the difference in the network usage.

6. Ages Indicators in Queries

A way to gain more insight on observed pedophile activity is to study the distribution of age indicators in queries [11]. Notice that age indicators are sometimes used in other contexts than pedophile activity, especially when parents seek content suitable for children of a certain age. However, one can observe on Figure 2 that ages indicators have similar behavior to those obtained for the pedophile group, and are therefore closely related to the topic.

We plot the distribution of age indicators on Figure 3: for each integer \( n \) lower than 17, we plot the number of queries of the form \( nyo \) in each dataset (\( yo \) stands for years old). The three plots have similar shape, with mostly increasing values from 1 to 10, a little drop at 11, a peak at 12 and a fall from 13 to 16. These values for KAD are below the values for the eDonkey servers, which is due to the fact that this dataset is a bit smaller than others and that pedophile queries are rarer in it. The key point here is that the distributions are very similar in all three datasets. This indicates that, although the amount of pedophile activity varies between systems, its nature is similar, at least regarding ages.

**Figure 3.** Distribution of age indicators in our three datasets.
7. Quantifying Pedophile Activity in KAD

In [4], the authors establish a method to quantify the fraction of pedophile queries in eDonkey. It relies on a tool able to accurately tag queries as pedophile or not, and on an estimate of the error rate of this tool. Such an approach cannot directly be applied to KAD though, as only a small (and biased) fraction of all queries may be observed in this system. We however show in this section how to derive the fraction of pedophile queries in KAD from the one in eDonkey.

In a given system, eDonkey or KAD here, we consider different sets of queries and we denote by \( Q \) the set of all queries, \( P \) the subset of pedophile queries in \( Q \), \( \bar{Q} \) the subset of queries composed of one word among the 72 monitored keywords, \( \bar{P} \) the subset of pedophile queries with one word, \textit{i.e.} consisting of one of the 19 monitored pedophile keywords (and so: \( \bar{P} = \bar{Q} \cap P \)). Figure 4 illustrates our notations.

In both our eDonkey measurements, \(|P|\) and \(|Q|\) may be directly estimated, as shown in [4], and one can then obtain the fraction \(|P|/|Q|\) of pedophile queries in the dataset. We give the results for our two measurements in Table 1. On the contrary, in KAD, one may only estimate \(|\bar{P}|\) and \(|\bar{Q}|\).

![Figure 4. The different sets of queries defined for each dataset.](image)

| Dataset | \( |P|/|Q| \) | \(|\bar{P}|\) | \(|\bar{Q}|\) | \( \alpha \) | \( \beta \) |
|---------|----------------|--------|--------|--------|--------|
| edonkeyFR | 2.554 \( \cdot 10^{-3} \) | 74,557  | 241,152 | 1.431 \( \cdot 10^{-3} \) | 0.2502 |
| edonkeyUA | 2.668 \( \cdot 10^{-3} \) | 46,763  | 166,154 | 1.538 \( \cdot 10^{-3} \) | 0.2251 |
| KAD     | n/a             | 30,821  | 250,000 | n/a    | n/a    |

However, we define \( \alpha = (|\bar{Q}| - |\bar{P}|)/|\bar{Q}| \) and \( \beta = |\bar{P}|/|\bar{P}| \), which capture the probability for a non pedophile query, respectively pedophile, to make a query of one word among one of our monitored keywords. Given the definition of \( \alpha \) and \( \beta \), there is no \textit{a priori} reason to assume that they have significantly different values between eDonkey and KAD. From the definitions of \( \alpha \) and \( \beta \), we have:
\[ \alpha = \frac{|Q| - |P|}{|Q| - |P|} \quad \Rightarrow \quad |Q| = \frac{\alpha |P| + |Q| - |P|}{\alpha} \]

\[ \beta = \frac{|P|}{|P|} \quad \Rightarrow \quad |P| = \frac{|P|}{\beta} \]

Then, the following expression holds:

\[ |P| = \frac{|P|}{\beta} \times \frac{\alpha}{\alpha |P| + |Q| - |P|} = \frac{\alpha |P|}{\beta |Q| + (\alpha - \beta) |P|} \tag{1} \]

We now use expression (1) to infer the fraction of paedophile queries that were submitted in the KAD P2P network during our experiment. Using the values from Table 1 and the average values of \( \alpha \) and \( \beta \) between our eDonkey datasets, we obtain:

\[ \frac{|P|}{|Q|} \approx 0.087\% \pm 0.008 \]

This value is of similar magnitude to the one of eDonkey (approx. 0.25%) but close to three times lower.

This estimation of \( \frac{|P|}{|Q|} \) relies on the value of \( \alpha \). One may wonder whether the choice of keywords from which we built \( Q \setminus P \) has a significant impact on the estimated value of \( \frac{|P|}{|Q|} \) in KAD. We check this as follows: we randomly select 1,000 subsets of 26 keywords out of the 53 keywords which compose the queries in \( Q \setminus P \). We then compute, for each subset, the number of queries consisting of exactly one of those keywords and the resulting value of alpha. For eDonkeyFR, we obtain an average value of \( \overline{\alpha} = 0.000889 \) (minimum: 0.000256, maximum: 0.00153, and 90% of the values in [0.000463;0.00133]). For eDonkeyUA, we obtain an average value of \( \overline{\alpha} = 0.00105 \) (minimum: 0.000352, maximum: 0.00172, and 90% of the values in [0.00062;0.00148]). This means that we would obtain very similar results with 26 keywords only and so we may be confident in our estimate obtained with 53 keywords.

8. Conclusions

We performed a comparative study of two large-scale peer-to-peer networks, KAD and eDonkey, with regards to the queries related to child pornography. We designed a methodology to collect and process datasets allowing to compare them in a relevant manner. We obtained the counter-intuitive result that pedophile keywords are significantly more present in eDonkey than in KAD, despite the higher anonymity level it provides. On the contrary, our study of age indicators in queries showed that the nature of pedophile queries is similar in these systems. We finally established the first estimate of the fraction of pedophile queries in KAD. We obtained a value close to 0.09%, which is of the same magnitude but significantly lower than in eDonkey (0.25%).

Our approach here is similar to the one used in [4]: we focus on search queries, which help to grasp the demand for pedophile material. It differs from [16,17] which focused on the files. In P2P networks such as eDonkey and KAD, a single file may have several names, most of which describe its content. However, filenames are prone to pollution and often exhibit keywords unrelated to the real content of
the file, for instance a pedophile file may have a non-pedophile name [23, 24]. Thus, estimations relying on specific filenames are likely to underestimate the true extent of child pornography distribution, while estimations relying on file-based honeypots are likely to overestimate the demand due to false-positive download requests. Query-based estimations using search requests do not suffer from such a bias, but, as mentioned earlier, may be impacted by repetitive queries from regular pedophile users. Nevertheless, both the considered P2P networks (KAD and eDonkey) should be equally affected, thus making their comparison valid to this regard.

Our contributions open various directions for future work. In particular, our methodology may be applied to compare other systems, and our datasets may be used to perform either deeper analyses on pedophile activity or on general search engine behaviors.

Acknowledgments

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

Author Contributions

RF participated in the study design, carried out analyses and interpreted the data, before writing the first manuscript in consultation with the co-authors. TC, IC and OF participated in the study design, data collection and experiment analysis. ML and CM participated in the study design and data analysis. ID was of significant help to collect data.

Conflicts of Interest

The authors declare no conflict of interest.

References


© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).