Content Pollution Quantification in Large P2P networks : a Measurement Study on KAD

Guillaume Montassier, Thibault Cholez, Guillaume Doyen, Rida Khatoun, Isabelle Chrisment*, Olivier Festor**
Université de Technologie de Troyes, STMR (UMR CNRS 6279), France
**INRIA Nancy-Grand Est / *LORIA - ESIAL, Nancy University, France
Email: {firstname.name}@utt.fr; *{firstname.name}@loria.fr

***Abstract*—Content pollution is one of the major issues affecting P2P file sharing networks. However, since early studies on FastTrack and Overnet, no recent investigation has reported its impact on current P2P networks. In this paper, we present a method and the supporting architecture to quantify the pollution of contents in the KAD network. We first collect information on many popular files shared in this network. Then, we propose a new way to detect content pollution by analyzing all filenames linked to a content with a metric based on the Tversky index and which gives very low error rates. By analyzing a large number of popular files, we show that 2/3 of the contents are polluted, one part by index poisoning but the majority by a new, more dangerous, form of pollution that we call index falsification.**

***Index Terms*—KAD; pollution of contents; pollution detection**

I. Introduction

P2P networks, and particularly Distributed Hash Tables (DHT), are today a major usage of Internet involving several millions of users. Their scalability, robustness and small infrastructure costs make them a perfect architecture to support file sharing applications, among others. However, the lack of central authority and malicious peer's behaviors can disturb the services offered by the P2P system. One of the major issues affecting large open P2P file sharing networks is the pollution of the shared contents. The pollution is defined as an irregular usage of the network to avoid a peer to access a desired content. Different forms of pollution have been identified affecting P2P networks by corrupting either the indexation of files [8] or files' content itself [9]. However, those studies were conducted several years ago on the FastTrack and Overnet P2P networks that are not used anymore today. Since 2005, no major investigation on P2P systems was done to quantify or describe the forms of pollution affecting current P2P networks.

We propose in this paper a measurement study of the pollution affecting the widely deployed KAD P2P network and we show that the pollution is widely diffused. In particular, we highlight and quantify a new form of pollution we call index falsification that aims at making a user access to a content totally unrelated to the desired one and potentially harmful.

This paper is structured as follows. Section II presents the related work on KAD and the different forms of pollution. Section III describes the index falsification pollution method and our strategy to detect it. Section IV presents the assessment of our pollution metric and our results on the quantification and the characterization of the pollution we obtained by investigating a large number of popular files in KAD. Finally, Section V concludes the paper and outlines our future work.

II. Related work

A. The KAD P2P network

KAD is a widely deployed P2P file sharing network implemented in the eMule client. As a file sharing application, KAD's Kademlia based DHT uses a double-indexation mechanism to index the shared files. The first level associates keywords with files while the second associates files with sources (peers sharing the file). Each KAD node has a random 128 bits "KADID" determining its position in the DHT and the references of which it is in charge of. When sharing a file, the raw data and all the keywords associated with its name are hashed separately with a MD4 function generating an ID which is then published into the DHT. Firstly, the file's information (fileID, filename, etc) are published towards the hash of each keyword (keywordID). Secondly, the peer publishes its own information (IP address, port, etc) towards the hash of the file (fileID) to be indexed as a potential source.

Several research papers have been written about KAD, covering many aspects of this P2P network. Some studies like [14] focused on large scale monitoring of peers participating in the network showing that KAD is mainly used in European countries and in China. [5] highlighted some flaws decreasing the routing efficiency of KAD and proposed some fixes. The security of KAD has also been widely investigated. The Sybil attack, which consists in the insertion of malicious nodes in the DHT, is particularly critical on this network [13] and some countermeasures have been proposed against it [1]. However, despite the great interest of the academic community for KAD, no study actually monitored the pollution of contents, which is one of the major issues of the network.

B. Pollution in P2P file sharing networks

Several forms of pollution exist and were studied in real P2P networks. The first form, called data pollution, consists in the sharing of files (decoys) whose content is deliberately damaged. The authors of [3] showed that the best strategy to spread such a pollution is to limit the number of polluted files advertised and to make them very popular, that is to say, shared by a lot of sources. The second form of pollution, called index poisoning or meta-data pollution, consists in corrupting the indexation mechanism of the P2P system by advertising

many fake files which are actually shared by no peer or by advertising unexisting sources. In 2004, Liang et al. were the first to study pollution [9] by analyzing different files related to 7 popular songs shared in the FastTrack network and showed that, for some popular songs, more than 50% of the versions were polluted. In 2005 [8], they considered 10 popular songs and investigated the pollution in both the FastTrack and the Overnet P2P networks showing that both were polluted and that Overnet is particularly affected by the index poisoning pollution. More recently, a study [10] proved that the index poisoning attack can also affect the KAD network by corrupting DHT entries, either by publishing fake records on the responsible peers or by inserting malicious nodes which are close to them.

Several solutions have been proposed to limit the pollution. Liang et al. [8] proposed a blacklisting scheme to avoid polluters but the list is provided and updated in a central way opposed to the P2P paradigm. Reputation systems [4] can help fighting the pollution but they introduce a significant overhead and are vulnerable to malicious votes among other attacks. Winnowing [12] proposed a collaborative filtering done by the indexing peers to limit the index poisoning but it is vulnerable to malicious nodes insertion. None of the proposed solutions is actually deployed so far because of the limitations.

III. Detecting KAD's DHT pollution

A. The index falsification pollution

We highlight and quantify a new form of pollution widely spread in the KAD network. While index poisoning advertises many unexisting files which can not be downloaded, index falsification pollution consists in advertising a single file under many different filenames, and consequently many different keywords, which are totally unrelated to the real content. Each wrong filename is artificially made popular by the polluters by exploiting a KAD weakness that allows a peer to manually publish the number of estimated sources for a given filename and which helps to make the pollution highly visible.

This form of pollution is the most harmful possible because it leads the user to download undesirable content, wasting the network resources in the process and because the downloaded file can be dangerous for the user safety. The real content can be a malware or a video hurting users' feelings (e.g. pornographic or paedophile contents). Besides, this pollution creates many false positives when monitoring illegal files within the KAD network. In fact, some users can be monitored accessing a file they did not even look for. While KAD's users can suffer from this pollution on a daily basis, no study so far investigated this problem.

B. Strategy for detecting the index falsification

The heart of index falsification is to attach many different names to a file. To detect this pollution, we must gather all the different filenames attached to a file and evaluate their consistency. However, KAD's double-indexation scheme makes this information hard to retrieve. From a keyword search, one can obtain the different files linked to the keyword and their details (filename, size, etc.). But all the different files collected from a keyword search include the desired keyword in their filename, for instance "avatar", and will appear consistent even if some files are indexed through other unrelated keywords in the DHT. From a source search, one can obtain all the sources of a file, no matter which filename is used by the sources. However, the filename is not an information published on the DHT at this level. Only the sources (IP address, port, etc.) are linked to the fileID. So, the diversity of filenames can not be obtained by regular DHT lookups, neither by a *keyword search* lookup nor by a *source search* lookup.

Filename: The.Big.Bang.Theory.4x09.The.Boyfriend.Complexity.ENG... FileID: C0F8BFA37E0DD0A4585CD3B90B9F4D26 Number of responding sources: 50	
Found filenames	#
The.Big.Bang.Theory.4x09.The.Boyfriend.Complexity.ENG.-.sub.FR...	30
The.Big.Bang.Theory.4x09.The.Boyfriend.Complexity.VOSTFR.HD...	12
The Big Bang Theory 4x09 The Boyfriendplexity Vostfr Hdtv Xvid...	3
The.Big.Bang.Theory.S04E09.VOSTFR.HDTV.XviD.avi	2
409 The.Big.Bang.Theory.4x09.The.Boyfriend.Complexity.VOSTFR...	1
The Big Bang Theory - 4x09 - VostFr.avi	1
The.Big.Bang.Theory.S04E09.VOSTFR.HDTV.XviD-TheOdusseus.avi	1

TABLE I: Example of consistent filenames retrieved from the responding sources for a clean file

Filename: Indiana Jones Et Les Aventuriers De L'Arche Perdue-Fr-Dvdrip... FileID: 7B9F403468CD821C38885E7777153C1C Number of responding sources: 175	
Found filenames	#
Xxx Marc Dorcel - Russian Institute Lesson 1 (Sex, Porno, Lesbian...	4
The Best Of The Doors.rar	2
[DIVX-ITA]-Disney Pixar-Wall-E-2008-Italian Ld Dvdrip Xvid...	1
[DIVX-ITA] The Twilight Saga New Moon.avi	1
Dexter Fr Saison 3.rar	1
Shrek.2.(Fr.DvdRipp).Teste.by.www.FreeDivx.org.avi	1
Smallville 6x10 Hidro [DVD+DVB][Spanish-English][by jesuscas]...	1
THE SOCIAL NETWORK [par emule island.com tp].avi	1
Windows 2003 Server.iso	1
...	...

TABLE II: Part of conflicting filenames retrieved from the responding sources for a polluted file

However, we can obtain the different filenames linked to a fileID at the beginning of the download process. When a file is selected for download after a keyword search, the potential sources are first retrieved thanks to the DHT. Then, a TCP connection is initiated toward each source to request the download. A function implemented in KAD clients, but external to the KAD protocol, allows a peer to get detailed information about the file directly from the responding sources and can show conflicting filenames in case of pollution. Concerning a clean file (table I), the majority (30) of the responding sources share the same filename and the others are clearly related to the desired content. On the opposite, concerning a polluted file (table II), the sources show a lot of totally different filenames, conflicting with each other and with the desired content. We used a modified aMule client in order to collect, for a given file, all the filenames advertised by the responding sources.

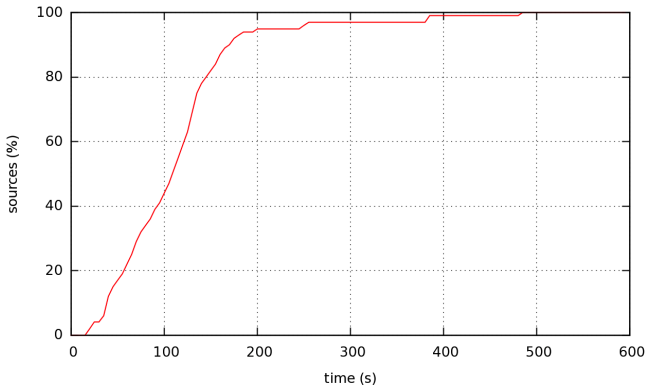

Fig. 1: % of the responding sources discovered in time

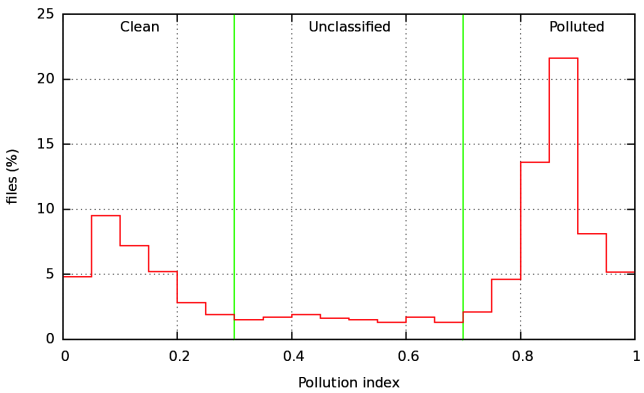

Fig. 2: Distribution of files according to the pollution index

C. Similarity metric for pollution detection

Given a fileID, we want to determine if the file is trustworthy or polluted by index falsification. Our detection is based on the overall consistency of the different filenames given by the sources. To compute the similarity of two filenames, we use a metric to evaluate the similarity between their set of keywords. Let X and Y be two sets of keywords, X being the keywords associated with the desired filename and Y being the keywords associated with a filename retrieved from a source. The Tversky index [15] is a generic similarity metric (when $\alpha = \beta$) used in data mining and defined by:

$$S(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha * |X - Y| + \beta * |Y - X|} \quad (1)$$

$S(X,Y) \in [0,1]$ and more precisely returns 1 if both filenames are the same and 0 if they have no common keyword. We define our pollution coefficient P for each file X as a function of the average of similarity coefficients computed for all filenames Y_i retrieved from the n sources.

$$P(X) = 1 - \frac{\sum_{i=1}^{n} S(X,Y_i)}{n} \quad (2)$$

IV. Content Pollution quantification in KAD

A. Investigating the shared contents

Collecting all the files shared within a P2P network is an impossible task. To quantify the pollution from a significative sample of the users' interest, we based our experiment on the top 100 of the most downloaded contents in 2010, according to one of the major BitTorrent indexation website[1] which receives more than one hundred million searchs a year. To have a most significative sample, for each content from top 100 we collected the twenty related files that show the highest number of estimated sources, resulting in 2000 files investigated. In fact, previous studies [2] and [7] showed that a higher number of sources estimated for a file increases its downloads.

Before collecting the different filenames for 2000 files, we wanted to know how fast the real sources of a file are found in order to define the duration of the experiment. Based on a sample of 150 files, we show that the number of responding sources quickly increases during the first 200s after the start of the download and is then stable (figure 1). We chose to wait up to 300s to collect the filenames, afterwards the number of missed sources is negligible. This result also indicates how fast the detection could be performed in real time.

B. Metric assessment

Every classification is inevitably prone to errors. In our case, some unpolluted files can be tagged as polluted (false positive) and inversely, polluted files can be tagged as unpolluted (false negative). To provide a reliable pollution detection, we need to define suitable parameters for the similarity metric and for the detection thresholds. To define these values and estimate the metric's error rates, we used a methodology close to [6] by asking a set of 10 experts to manually evaluate the pollution for a sample (20%) of the files we collected. By analyzing the different filenames, they tag files as *polluted*, *clean* or *unclassified* through a web interface showing the associated filenames as presented in tables I and II.

We tried some particular forms of the Tversky similarity metric and found that the best detection results according to the expert votes were given for $\alpha = \beta = 0.5$ which is known as the Dice coefficient [11] and can also be written like:

$$S(X,Y) = \frac{2 * |X \cap Y|}{|X| + |Y|} \quad (3)$$

Then, we set the detection thresholds to best match the expert votes: files for which $P(X)$ is under 0.3 are tagged as clean while those above 0.7 are polluted. Concerning the few files between 0.3 and 0.7, we consider that the metric can't reliably determine their state. Figure 2 displays the distribution of sources according to our metric. It appears that we have a bimodal distribution with peaks at 0.1 and 0.9 showing that our metric naturally creates two major classes for our data. As a result, our metric based on the Dice coefficient with these detection thresholds gives low error rates: **3.5%** of false positive and **0.8%** of false negative.

We investigated the false positives of our metric and found that they are due to movies presenting both an original

[1] http://torrentfreak.com/bittorrent-zeitgeist-what-people-searched-for-in-2010-101227/

and a localized title. For example, we downloaded a file called *"el Cigno Nero Sub Ita.avi"* and the name for the majority of sources was *"Black.Swan.2010.DVDSCR.XviD-TiMKY.avi"*. From the metric point of view, these two names do not appear related, only a semantic knowledge of the translated words or of the different localized names set for a movie can lead to the right classification given by the experts.

C. Quantification and characterization of KAD's pollution

To quantify the index falsification, we applied our final metric on the popular files we investigated. Table III shows the final classification of these files. First we can see that despite a large number of announced sources, no responding sources can be found for 20.5% of the files which clearly indicates an index poisoning pollution. However, the major form of pollution is the index falsification affecting 41.1% of the considered KAD's files. These two forms of pollution represent 61.5% of the 2000 popular files, leaving less than a third (28.6%) clean. According to the expert votes 78% of the unclassified files are clean and 12% polluted.

Type	quantification (%)
No responding source (index poisoning)	20.5
Polluted (index falsification)	41.1
Clean	28.6
Unclassified	8.6

TABLE III: Global pollution quantification

Finally, we analyzed the number of corrupted files for each entry of the top 100. It appears that all the entries of the top 100 were concerned by this pollution, from the less infected: *The big bang theory* with 25% of polluted files out of the 20 considered, to the most: *Avatar* with all its top 20 files polluted. Besides, it appears that the index falsification pollution does not only target copyrighted content since the entry *Ubuntu* had fifteen polluted files out of twenty.

Moreover, we investigated more precisely the different filenames of polluted files to determine by which type of content they might be polluted with. We used two lists of keywords, one related to paedophile contents [6] and the other to pornographic contents. We then searched for those keywords among the filenames advertised for our 41.1% of files affected by index falsification. Table IV shows the resulting characterization: 8.8% of them are referenced by at least a paedophile filename and more than 55% by a pornographic name.

Contents	quantification (%)
Child pornography	8,8%
Pornography	55,7%
Other	35,3%

TABLE IV: Type of contents found in index falsification

V. Conclusion

We presented in this paper a new type of pollution widely spread in the KAD network and called index falsification. We proposed a very efficient metric to detect it, assessed by the evaluation of experts. We applied our metric to a large number of popular contents to quantify the pollution and found that popular files are highly infected by pollution, with more than 41% of files infected by index falsification and more than 20% by index poisoning. Moreover, our results show that index falsification is an intentional and harmful pollution, linking undesirable contents (potentially paedophile and pornographic) through regular keywords and sometimes even related to cartoons.

In our future works, we will investigate the polluting behaviors in order to understand precisely how this pollution is achieved. Then, we will design a detection mechanism which can operate earlier in the download process to avoid the initialization of many connections towards the responding sources. Our solution will also need to be suitable for real implementations (by keeping backward compatibility and minimizing the overhead) in order to protect current P2P networks.

References

[1] Thibault Cholez, Isabelle Chrisment, and Olivier Festor. Efficient DHT attack mitigation through peers' ID distribution. In *Seventh International Workshop on Hot Topics in Peer-to-Peer Systems - HotP2P 2010*, Atlanta USA, 04 2010. IEEE International Parallel & Distributed Processing Symposium.

[2] Thibault Cholez, Isabelle Chrisment, and Olivier Festor. Monitoring and Controlling Content Access in KAD. In *International Conference on Communications - ICC 2010*, Capetown South Africa, 05 2010. IEEE.

[3] Nicolas Christin, Andreas S. Weigend, and John Chuang. Content availability, pollution and poisoning in file sharing peer-to-peer networks. In *EC '05: Proceedings of the 6th ACM conference on Electronic commerce*, pages 68–77, New York, NY, USA, 2005. ACM.

[4] Cristiano Costa and Jussara Almeida. Reputation systems for fighting pollution in peer-to-peer file sharing systems. In *P2P '07: Proceedings of the Seventh IEEE International Conference on Peer-to-Peer Computing*, pages 53–60, Washington, DC, USA, 2007. IEEE Computer Society.

[5] Hun Jeong Kang, Eric Chan-Tin, Nicholas Hopper, and Yongdae Kim. Why kad lookup fails. In *Peer-to-Peer Computing 09*, pages 121–130, Atlanta, USA, 09 2009. IEEE.

[6] Matthieu Latapy, Clémence Magnien, and Raphaël Fournier. Quantifying paedophile queries in a large P2P system. In *IEEE International Conference on Computer Communications (IEEE INFOCOM 2011) Mini-Conference*, 2011.

[7] Uichin Lee, Min Choiz, Junghoo Choy, M. Y. Sanadidiy, and Mario Gerla. Understanding pollution dynamics in P2P file sharing. In *5th International Workshop on Peer-to-Peer Systems (IPTPS'06)*, Santa Babara, CA, USA, February 2006.

[8] J. Liang, N. Naoumov, and K. W. Ross. The Index Poisoning Attack in P2P File Sharing Systems. In *INFOCOM 2006. 25th IEEE International Conference on Computer Communications*, pages 1–12. IEEE, 2006.

[9] Jian Liang, Rakesh Kumar, Yonjian Xi, and Keith W Ross. Pollution in p2p file sharing systems. In *INFOCOM*, pages 1174–1185. IEEE, 2005.

[10] Thomas Locher, David Mysicka, Stefan Schmid, and Roger Wattenhofer. Poisoning the Kad Network. In *11th International Conference on Distributed Computing and Networking*, Kolkata, India, 01 2010.

[11] Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999.

[12] Kyuyong Shin, Douglas S. Reeves, Injong Rhee, and Yoonki Song. WINNOWING : Protecting P2P Systems Against Pollution By Cooperative Index Filtering. Tech report TR-2009-2, Department of Computer Science, North Carolina State University, 2009.

[13] Moritz Steiner, Taoufik En-Najjary, and Ernst W. Biersack. Exploiting kad: possible uses and misuses. *SIGCOMM Comput. Commun. Rev.*, 37(5):65–70, 2007.

[14] Moritz Steiner, Taoufik En-Najjary, and Ernst W Biersack. A global view of kad. In *IMC 2007, ACM SIGCOMM Internet Measurement Conference, October 23-26, 2007, San Diego, USA*, 10 2007.

[15] Amos Tversky. Features of similarity. In *Psychological Review*, volume 84, pages 327–352, 1977.