Forensics for Detecting P2P Network Originated MP3 Files on the User Device

Heikki Kokkinen and Janne Nöyränen

Nokia Research Center,
Itämerenkatu 11-13, 00180 Helsinki, Finland
{heikki.kokkinen,janne.noyranen}@nokia.com

Abstract. This paper presents how to detect MP3 files that have been downloaded from peer-to-peer networks to a user hard disk. The technology can be used for forensics of copyright infringements related to peer-to-peer file sharing, and for copyright payment services. We selected 23 indicators, which show peer-to-peer history for a MP3 file. We developed software to record the indicator values. A group of selected examinees ran the software on their hard disks. We analyzed the experimental results, and evaluated the indicators. We found out that the performance of the indicators varies from user to user. We were able to find a few good indicators, for example related to the number of MP3 files in one directory.

Keywords: Peer-to-peer, P2P, MP3, forensics, binary classification, legal, copyright.

1 Introduction

This paper discusses technology to detect which Motion Picture Expert Group Audio layer 3 (MP3) files on the user device originate from peer-to-peer (P2P) networks. P2P file sharing applications and networks include for example Napster, Kazaa, Gnutella, eDonkey, and BitTorrent. P2P file sharing has created most of the traffic in the Internet in the past years. A significant amount of this traffic is copyright content with licenses, which do not allow sharing in the P2P networks. Though peer-to-peer networks are infamous for copyright infringements, there are also many legal ways to use P2P file sharing. Napster P2P application was enhanced with models to pay for the content [1]. The rights owners may allow the P2P file sharing with Creative Commons licenses [2] or in other ways. Increasing amount of companies use P2P file sharing to decrease their Content Distribution Network (CDN) costs, like Blizzard with World of Warcraft [3]. In a recently published post-payment copyright service the users are able to legalize their unauthorized media content by paying the copyright fees after downloading [4]. This paper describes the technology, which supports the post-payment copyright system by helping the user to select the files for which he wants to purchase the post-payment licenses. The technology suits also well for forensics purposes in finding evidence for copyright infringements. It is important to notice that post-payment copyright system and forensics are two different use cases for the technology, and they should not be mixed together.

M. Sorell (Ed.): e-Forensics 2009, LNICST 8, pp. 10 – 18, 2009.

The attempts to detect copyright content in the P2P networks have often been related to investigations of copyright infringements. Broucek et. al. describe general methodology for digital evidence acquisition for computer misuse and e-crime [5]. The ISPs have the best capabilities to collect information about the behavior of the investigated users. This kind of network collection is the most commonly used method for P2P copyright infringement forensics at the moment. Generic P2P traffic detection and prevention have been discussed in [6], and with emphasis on traffic mining in [7].

A commonly proposed method to detect copyright infringements in the user device is watermarking [8]. Koso et. al. apply Digital Signatures to watermarking [9]. The watermarking is a technology to embed information to content so that it does not alter the human perception of the content and so that the information is difficult to remove. The watermark is at investigation time used to track the source of the content. Digital Time Warping achieves independence from encoding and sampling [10]. An option to evaluate the source of a MP3 file is to carry out MP3 encoder analysis [11]. An application called Fake MP3 detector differentiates the files, which have different content as the name suggests [12]. The copyright infringement detecting and tracing are studied in [13].

In this paper we use an empirical method to detect which MP3 files on the user device originate from P2P networks. We identify 23 indicators, which show that a MP3 file has been downloaded from P2P network. We let six examinees to run the research software on their hard disks. All examinees have files originating from P2P network, and most of them have self-ripped files, as well.

After running the software the users manually classify which files are originated from P2P network. The research software records the values of all indicators for each MP3 file. We use the sensitivity and specificity performance metrics, which have commonly been used in binary classification context. The results show that the most suitable indicators vary from person to person, but a few indicators reveal well the P2P download origin.

In addition to the forensics use, the main application of the results is to help a user to select, which MP3 files are authorized and to which ones the user should purchase the license using the post-payment copyright system or by other means. The studied method evaluates the indicators. The P2P origin is in many cases a rule of thumb differentiating the authorized and unauthorized files of a typical user in Finland. Nevertheless, not nearly all P2P files are illegal, neither nearly all MP3 files without P2P history are legal. Even if the indicators were able to differentiate the P2P originated files with 100% accuracy, the legal status of the studied MP3 files would remain inaccurate. On the other hand, if the user was expected to classify his files to legal and illegal fully manual, going through thousands of files would be tedious, and this technology provides for the user a great help for the selection.

2 Materials and Methods

In this study we selected 23 indicators, which potentially show that a MP3 file is originated from a P2P network. We had six examinees. We developed software, which can run on an examinee's PCs and recorded the results of the indicators for

each MP3 file. The examinees ran the software and classified the origin of the files. We used three types of indicators: file specific indicators, directory specific indicators, and album specific indicators.

2.1 File Indicators

The file indicators try to classify the files, in this case MP3 tracks, individually.

1) *The file name, file path or file contains a P2P sharing group name like "EiTheLMP3".*The list of names was collected from two sites: [14] and [15].
2) *The file path contains 1337 speak like "m@ke".*
3) *ID3 tag comment field has an URL address like" http://www.torrentreactor.net/".*
4) *ID3 tag comment field contains 1337 speak.*
5) *ID3 tag title or comment field has a tag of a ware group tag like "RAGEMP3".*
6) *ID3 tag comment field is not empty.*

2.2 Directory Indicators

The directory indicators go through the files in a directory and compare them with each other.

7) *The file path has any of the following words: download, or shared.*
8) *A directory contains over 40 MP3 files.*
9) *A directory contains over 25 MP3 files.*
10) *The music in the directory has a longer total duration than 80 min.*
11) *The MP3 directory contains more than 3 other than music files.*
12) *The directory contains a file with the following type .nfo.*
13) *The directory contains a file with a following type .url, .torrent or .info.*
14) *There is a .txt file the same directory*
15) *There are no other tracks from the same album according to the album ID3 tag.*

2.3 Album Indicators

The album indicators study the common characteristics of the files, which have the same album ID3 tag.

16) *The track number is filled in some, but not in all tracks of the album.*
17) *All tracks are not encoded the same way (VBR or CBR)*
18) *The album files have different bitrate, only used for CBR.*
19) *All tracks do not have the same sampling rate.*
20) *Tracks vary from mono to stereo.*
21) *Many file indicators are present for the tracks of the album.*
22) *The file names contain capital and non-capital letters in a varying way.*
23) *The file names contain symbol characters in a varying way.*

2.4 Examinees

The examinees were selected so that they had a large amount of files, which originated both from P2P networks and from personal ripping from Compact Discs

(CD). The table 1 describes the MP3 software of the examinees. For simplicity the source of the MP3 files was expected to be either P2P network or personal ripping of CDs. As background information about the source we collected the users' CD ripper and MP3 encoder, and P2P file sharing application. MP3 re-tagging alternates the possibility to carry out the detection with the selected indicators. In some cases the player may also change the files or directories.

Table 1. Examinees' MP3 related software

User	Ripper	Tagger	P2P	Player
1	EAC-LAME, iTunes	Tag-Scanner	Azureus, eDonkey	iTunes, WMP
2	Audio-grabber		Bittorrent, Limewire	WinAmp, Rythmbox
3	WMP		WinMX	WinAmp
4	WMP		WinAmp	-
5	Audio-grabber		DC++, Bittorrent	WinAmp
6				WinAmp

In the table 2 we characterize the users according to the number of studied MP3 tracks and the percentage of illegal files.

Table 2. The number of tracks and percentage of illegal files for a user

User	1	2	3	4	5	6
Tracks	3394	1511	1946	905	2017	811
Illegal%	16.8	93.1	99.2	12.3	82.2	100

2.5 Metrics for Indicator Characterization

The commonly used performance metrics for binary classification are sensitivity and specificity. The typical application of the binary classification is to use medical examinations to find out if the patient has a certain disease or not. The examination results are divided to true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*).

The sensitivity is defined

$$Sensitivity = \frac{TP}{TP + FN}. \tag{1}$$

The sensitivity describes which portion of the illegal files the indicator was able to find. The specificity is defined

$$Specificity = \frac{TN}{TN + FP}. \tag{2}$$

The specificity characterizes which part of the files, which were classified as legal, were really legal. When other than binary indicators are used or several binary indicators are combined together, it is possible to adjust the decision limit. If we want

to change the decision limit so that our sensitivity increases, we lose in specificity and vice versa.

3 Results

We calculate the sensitivity and specificity values for each examinee per indicator. The summary of the sensitivity and specificity analysis can be found in the figure 1. The indicators are sorted according to the sensitivity average, the best indicator is on the left and the worse on the right. The standard deviation error bars show that there is a large variation in the capability to indicate P2P history for a MP3 file in the indicators. In most use cases the specificity value should stay close to 100%. The average specificity of indicator 6 is below 40%, but as we can see later in this section, it works well with the data of a few examinees.

The best average indicator for this group of examinees was *10) The music in the directory has a longer total duration than 80 min*. It has close to 100% specificity and the highest sensitivity (around 30%). The following indicators have also a reasonable sensitivity and close to 100% specificity: *9) A directory contains over 25 MP3 files*, *8) A directory contains over 40 MP3 files, 16) The track number is filled in some, but not in all tracks of the album, 3) ID3 tag comment field has a URL, 17) All tracks are not encoded the same way (VBR or CBR), and 19) All tracks do not have the same sampling rate*. In the Figures 2, 3 and 4 we show the specificity and sensitivity of three individual examinees' data.

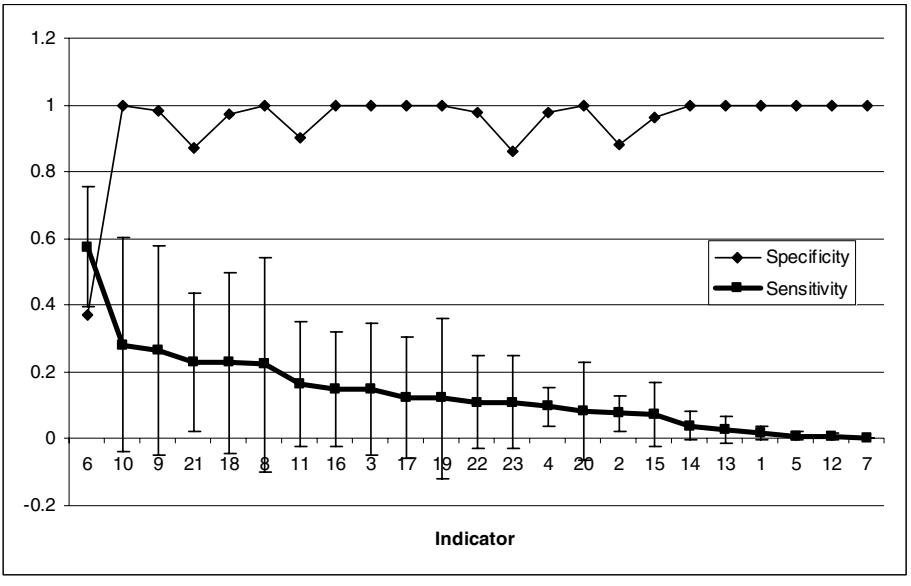

Fig. 1. Sensitivity average with standard deviation error bars and Specificity average

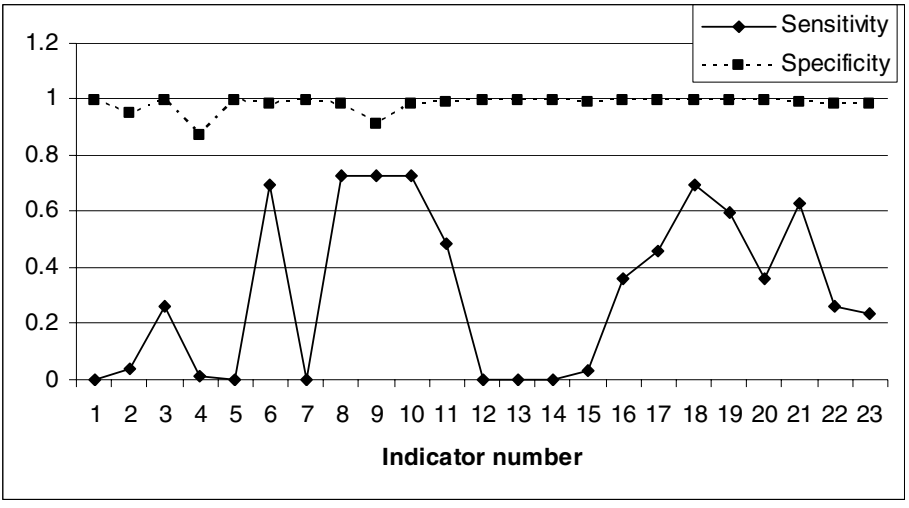

Fig. 2. Example of quite high sensitivity specificity (dashed)

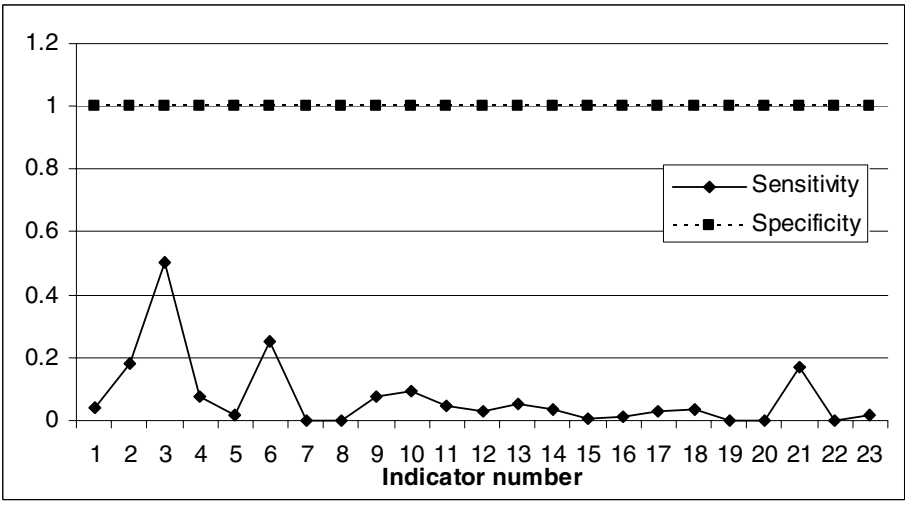

Fig. 3. Example of very high specificity (dashed) and low sensitivity

The figure 2 is a case where many indicators show that a file has been downloaded from P2P network and the specificity remains under control. Especially indicators *8) A directory contains over 40 MP3 files, 9) A directory contains over 25 MP3 files*, and *10) The music in the directory has a longer total duration than 80 min* perform well. These three indicators are related to each other and this examinee has downloaded many files one by one rather than as a whole album from P2P networks.

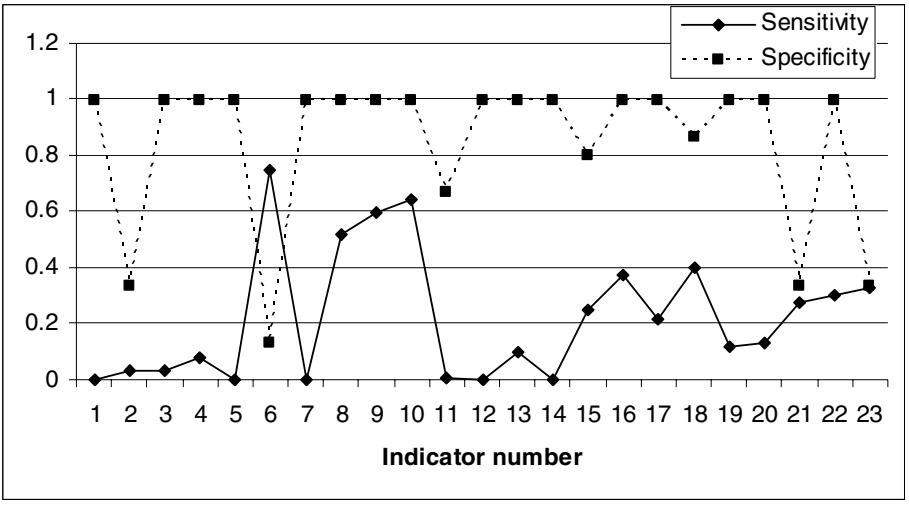

Fig. 4. Example of low specificity (dashed) and quite high sensitivity

The tracks are stored in one directory. Also indicators *18) The album files have different bit rate*, and *21) Many file indicators are present for the tracks of the album* have high sensitivity. The indicator *6) ID3 tag comment field is not empty* can keep the high specificity in this case.

In the figure 3 the specificity is constantly 100%. The best indicators are *3) ID3 tag comment field has a URL address*, *6) ID3 tag comment field is not empty* and *21) Many file indicators are present for the tracks of the album*. The high specificity value is obvious result in this case, because this examinee had 100% of the files from P2P networks.

In the figure 4 the challenge are the low specificity values with many indicators. Especially, the indicator 6 shows very low specificity. The generally best performing indicators are the group of 8, 9 and 10, which indicate a large number of MP3 files in one directory. Also indicator *16) The track number is filled in some, but not in all tracks of the album* has high sensitivity and close to 100% specificity.

4 Discussion

In this paper we studied the performance of 23 indicators, which show that a MP3 file is potentially from a P2P network. We evaluated the indicators with binary classification performance metrics: sensitivity and specificity. The best indicator was *10) The music in the directory has a longer total duration than 80 min achieved* close to 30% average sensitivity and practically 100% specificity. Generally the related indicators 8, 9 and 10 which indicate a large number of MP3 files in the same directory performed well. The number of files in a directory does not principally have anything to do with P2P networks. It is just a way how users organize their MP3 files in the directories. By using the number of files, we take a bold assumption that there

are only two main sources of MP3s, either ripping the CDs or downloading files from P2P network. P2P file sharing is so huge phenomenon that this assumption works especially with the people who either are investigated in the forensics methods or who are interested to use post-payment copyright type of services.

The obvious high specificity indicators *1) The file name, file path or file contains a P2P sharing group name like " EiTheLMP3"*, *2) The directory contains a file with the following type .nfo, .url, .torrent or .info, 3) ID3 tag comment field has a URL address like http://www.torrentreactor.net/", 4) The file path or file contains 1337 speak like "m@ke" or 5) ID3 tag title or comment field has a tag of a ware group like "RAGEMP3"* were not strongly visible in this group of examinees. The most common indicator of these was *3* showing URL address existence in the ID3 tag. It achieved an average of 15% sensitivity and practically 100% specificity. The related indicator *6* revealing any text in the comment had the highest sensitivity of all indicators, but due to very low specificity with a few users its accuracy was dropped significantly.

The specificity of the indicators varied from user to user significantly. One clear reason for very good specificity values was that a couple of examinees had practically all files from P2P networks. The number of examinees was rather small (6), and a few users did not have many files without P2P origin, making the specificity analysis less meaningful.

The results of this research can be used for forensics purposes to find out the P2P network origin of files on the device of the examined user. They can also be applied for post-payment copyright system to help the user to select the unauthorized MP3 files for license purchase. The examinees did not try to cover the origin of the P2P networks in their files. If one would systematically try to cover the traces by renaming, retagging and rearranging the files, these indicators may lose their effectiveness.

It would be interesting to research methods and algorithms, which could achieve the combinatory performance of all used indicators, and to study the performance of the methods by comparing to the performance of individual indicators. The studied indicators can individually be used to reveal the P2P origin of the MP3 files, if the examinees have not tried to remove the traces beforehand.

References

1. Alves, K., Michael, K.: The Rise and Fall of Digital Music Distribution Services: a Cross-Case Comparison of MP3.com, Napster and Kazaa. In: Cerpa, N., Bro, P. (eds.) Building Society Through E-Commerce, 1st edn., University of Talca, Talca (2005)
2. Creative Commons licenses, http://creativecommons.org/licenses/
3. World of Warcraft – Frequently Asked Questions, `http://www.blizzard.co.uk/wow/faq/bittorrent.shtml`
4. Kokkinen, H., Ekberg, J.E.: Post-payment copyright for digital content. In: 5th Consumer Communications and Networking Conference, CCNC, pp. 1278–1283. IEEE, Las Vegas (2008)
5. Broucek, V., Turner, P.: Computer Incident Investigations: e-forensic Insights on Evidence Acquisition. In: 13th Annual EICAR Conference, Grand-Duche du Luxembourg (2004)

6. Ho, G.L., Taek, Y.N., Jong S.J.: The method of P2P traffic detecting for P2P harmful contents prevention. In: 7th International Conference on Advanced Communication Technology, vol. 2, pp. 777–780 (2005)
7. Togawa, S., Kanenishi, K., Yano, Y.: Peer-to-Peer File Sharing Communication Detection System Using Network Traffic Mining. HCI (8), 769–778 (2007)
8. Nikolaidis, N., Giannoula, A.: Robust Zero-Bit and Multi-Bit Audio Watermarking Using Correlation Detection and Chaotic. In: Digital Audio Watermarking Techniques and Technologies: Applications and Benchmarks. Idea Group Inc. (IGI) (2007)
9. Koso, A., Turi, A., Obimbo, C.: Embedding Digital Signatures in MP3s. IMSA. pp. 271–274 (2005)
10. Sung, B., Jung, M., Ham, J., Kim, J., Ko, I.: Feature Based Same Audio Perception method for Filtering of Illegal Music Contents. In: 10th Int. conference on Advanced Communication Technology, ICACT, pp. 2194–2198 (2008)
11. Böhme, R., Westfeld, A.: Statistical characterisation of MP3 encoders for steganalysis. In: International Multimedia conference, workshop on Multimedia and security, pp. 25–34. Magdeburg, Germany (2004)
12. Fake MP3 detector, `http://www.sharewareconnection.com/fake-MP3-detector.htm`
13. Mee, J., Watters, P.A.: Detecting and Tracing Copyright Infringements in P2P Networks. In: International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies (ICNICONSMCL 2006), p. 60 (2006)
14. MP3 Kingz, `http://www.mp3kingz.org/`
15. NfoDB.com, `http://www.nfodb.com/section_4_mp3_nfo.html`