BTM – An Automated Rule-based BT Monitoring System for Piracy Detection

K.P. Chow, K.Y. Cheng, L.Y. Man, Pierre K.Y. Lai, Lucas C.K. Hui,
C.F. Chong, K.H. Pun, W.W. Tsang, H.W. Chan, S.M. Yiu
The University of Hong Kong
{chow,kycheng2,lyman,kylai,hui,chong,pun,tsang,hwchan,smyiu}@cs.hku.hk

Abstract

With the advent of peer-to-peer communication technologies, individuals can easily connect to one another over Internet for file sharing and online chatting. Although these technologies provide wonderful platforms for users to share their digital materials, its illegitimate use on unauthorized sharing of copyrighted files is increasingly rampant. With the BitTorrent (BT) technology, the tracking down of these illegal activities is even more difficult as the downloaders can also act as the distributors and cooperate to provide different parts of the same file for sharing. It is close to impossible for law enforcement agencies to trace these distributed and short-duration Internet piracy activities with limited resources. In this paper, we present the first automated rule-based software system, the BitTorrent Monitoring (BTM) System, for monitoring, recording, and analyzing suspicious BT traffic on the Internet. From a preliminary experiment on a real case, the system successfully located 126 distributors (a.k.a. seeders) for some Cantonese pop songs within 90 minutes.

1. Introduction

Since, the advent of the very famous Napster in 1999, peer-to-peer (P2P) applications have attracted a large user community which has been growing rapidly as never before. An Internet traffic management firm has estimated that 50-65 percent of all Internet downloads traffic, and 75-90 percent of all Internet uploads traffic, are P2P-related [6]. Without doubt, P2P technology offers a wonderful platform for individuals and organizations to share their digital materials worldwide extremely easily. Unfortunately, its illegitimate use on unauthorised sharing of copyrighted files is increasingly rampant and is reaching an alarming height. According to a third-party research, potential losses to the recording industry from P2P file-sharing was estimated at US$2.1 billion in 2004 [7]. More recently, a survey conducted in September 2005 indicated that the average number of files available for download at any given moment on P2P networks worldwide was nearly 2.8 billion [2]. In Hong Kong, according to some statistics from the Government [5], it is also found that the percentage of respondents who admitted that they would illegally download and upload files for sharing has doubled. It seems that the problem of Internet piracy is getting serious.

Among the few successful P2P protocols in existence, BitTorrent (BT) has evolved into the most popular networks [1] and has managed to attract millions of users since inception. By the end of 2004, BitTorrent was accounting for as much as 50% of all P2P-related traffic [3]. It wins user support through providing its users carefree sharing of large files at high downloading speed, while at the same time the requirements on bandwidth and hardware are remarkably small.

With the existence of the overwhelming private BitTorrent networks, it is difficult to gauge the actual numbers of BT users. What we are certain, however, is the tremendous loss to the media industries [9]. Over the years, law enforcement agencies have set out operations to fight against these illegal activities. With much of their effort, the world's first conviction of piracy of BitTorrent user was sentenced in the fall of 2005. But it is sad to say the outcome did not prove an effective deterrent to average BT users. Although many individuals realize that what they are doing is a kind of online piracy and is illegal under recently enacted legislation, they still pursue the file sharing as before. The thing that put their mind at ease is that the limited manpower available to law enforcement agencies is exceedingly insufficient for cracking down every single member of the enormous BT user base [8].

So, there is an imperative need to develop an automated system for monitoring these increasingly rampant BT activities.

The rest of this paper is organized as follows. The working principle of BitTorrent Technology is presented in Section 2. Section 3 gives an overview of the BTM system. Section 4 describes how BTM locates torrent files. Torrent analysis is discussed in details in Section 5. Finally, we discuss the result of a preliminary evaluation of the system and conclude the paper in Section 6.

2. BitTorrent (BT) basics

BitTorrent is a peer-to-peer file distribution protocol, which is designed to allow efficient distribution of large amounts of data and is one of the best ways to distribute large files, such as videos and high-quality MP3s, with minimal demand on hardware resources and communications bandwidth. BitTorrent uses a peer to peer strategy in which every user's computer contributes. Its idea is to redistribute the cost of upload to downloaders: when several people are downloading the same file at the same time, they upload pieces of the file to each other. Figure 1 shows the working protocol of a BitTorrent network with four peers.

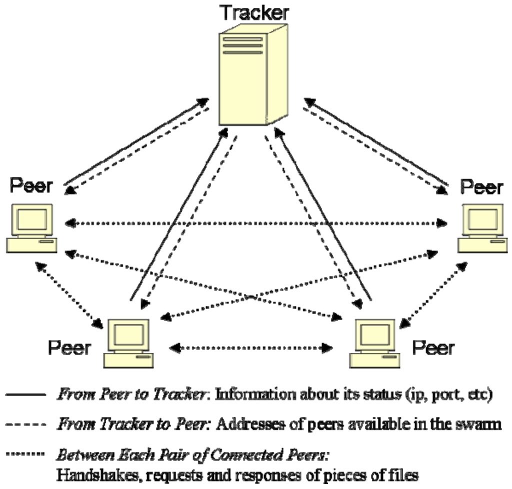

Tracker
Peer
Peer
Peer
Peer
From Peer to Tracker: Information about its status (ip, port, etc)
From Tracker to Peer: Addresses of peers available in the swarm
Between Each Pair of Connected Peers:
Handshakes, requests and responses of pieces of files

Figure 1. The working protocol of a BT network in a swarm of 4

A BitTorrent network is made up of four types of entities:

- *Tracker* is a server that coordinates the distribution of files. It acts as an information exchange center from which peers obtain necessary information about other peers to which they can connect.
- *Torrent file* is a file which contains metadata about the files to be shared. It contains the address of the tracker that coordinates communication between peers.
- *Peer* is a computer that participates in a download. Together, all peers (including seeders) sharing a torrent, are called a *swarm*.
- *Seeder* is a peer that has a complete copy of the file and offers it for download.

2.1. Making and publishing torrent file

To share a file using BitTorrent, the file owner generates a torrent file which contains metadata about the file to be shared and the URL of the tracker. As the file will normally be distributed in pieces, the torrent file also specifies the piece length used and a hash code for each piece so that other peers can verify the integrity of the pieces. After the torrent file is created, it is registered with a tracker and the file owner has to make the torrent file available to other Internet users by placing it on a website or elsewhere. The computer with the initial copy of the file is usually referred to as the initial seeder.

2.2. Obtaining torrent file and sharing files

After downloading the torrent file from the Internet, a user opens it with a BitTorrent client program, which is responsible for connecting to the tracker and managing the transfer of the file. The BitTorrent protocol splits files into a number of small pieces of pre-specified length. To ensure these pieces are error-free, they are checked by the client program using the hash algorithm SHA-1 [10] on their arrival. When a file is initially shared, peers do not have complete pieces to share with other peers. They need to connect directly to the initial seeder and begin to request pieces. As more peers appear, they begin sharing pieces with one another, instead of downloading directly from the seeder. Clients report information to the tracker periodically and in exchange receive information about other clients to which they can connect. It is worthy to note that the tracker is not directly involved in the data transfer throughout the course of file sharing. Special mechanisms have been developed for optimizing transfer rates, which are out of the scope of this paper and will not be discussed here.

3. Overview of the BTM system

An overview of the BTM System is depicted in Figure 2. To locate torrent files, BTM searches target websites specified by user-inputted URLs. It downloads and archives the located torrent files and starts its preliminary analysis on them. From the torrent files, BTM is able to connect to tracker(s) and retrieve the lists of peers currently participating in the torrent. It continues to connect to peers and gather data from each of them using the Peer Wire Protocol. Based on the information extracted from the torrent files, the trackers' responses and peers' statuses, BTM performs in-depth analysis and generate summary reports. The peer information gathered using the Peer Wire Protocol will be used by the rule engine to determine if any necessary action needs to be triggered.

Basically BTM consists of two modules, torrent file search and torrent analysis, implemented in software components Torrent Searcher and Torrent Analyzer, respectively. Torrent searcher is responsible for searching of torrent files on the websites specified by the user. Subsequently, the downloaded torrent files will be passed on to Torrent Analyzer. It gets in touch with the trackers and the peers and gathers useful information, on which the analysis is performed. Torrent Searcher and Torrent Analyzer will be described in more detail below.

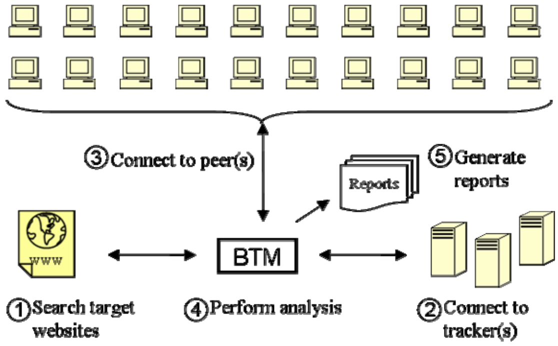


Figure 2. An overview of the BTM System

4. Torrent searcher

To launch a download with BitTorrent, users must have a torrent file on hand. BitTorrent itself is a file-sharing protocol only, as such, it relies on other mechanisms for the location of torrent files. Usually torrent files are made available to other Internet users through various channels. For example, indexing sites and public forums provide a convenient platform for BT users to exchange torrent files. In its preliminary stage, BTM targets public forums, where actual BT users communicate with each other and locate torrent files, come into contact with the peers, and set out its operation.

4.1. Searching Algorithm

BTM supports automated torrent searching by seeking torrent files on websites and exploring the hyperlinks over there. On each page directed by the hyperlinks, it continues to seek torrent files until a predefined level has been reached. A level with depth equals to three means that Torrent Searcher will catch all pages specified by the inputted URLs, plus all those that can be accessed in two clicks on any link from there. The concept of levels is illustrated in Figure 3.

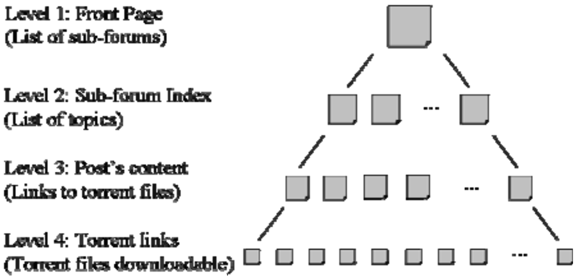


Figure 3. An illustration of the concept of levels

Depth first search (DFS) with finite depth (maximum length of search path) is adopted in BTM to maximize the number of torrent files located within a predefined stopping time. Figure 4 gives an illustration of the searching algorithm, where the numbered lines show the order of processing.

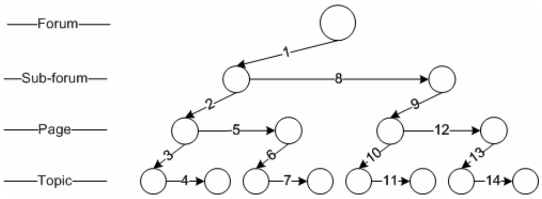


Figure 4. An illustration of depth first search algorithm

Link redirection

It is increasingly common to maintain torrent files in external servers, where they are typically stored in a form that requires a click to launch the download. The click action can be simulated by submitting a POST message to the server. BTM has been tested to work with many popular external servers.

4.2. Cookies Handling

To gain access to these public forums, a user needs to log on with a registered user login name and password. In BTM, the login process is simplified through the use of cookies, which are text files stored on clients' browsers, enabling servers to prepare customized web pages and track users' actions. As an initial step, a BTM user manually logs on the website with a registered user login and password. This is performed once for many subsequent auto-accesses. With cookies enabled on the website, Torrent Searcher automatically complete the login process and start seeking torrent files without user's intervention.

4.3. Keyword Searching

Public forums provide BT users with a convenient platform for open discussions. To share files, BT users usually supply some descriptions and comments together with the torrent files. This information may be useful for law practitioners to carry out their investigation. Multifarious contents may, however, appear in the links of a forum. BTM allows its users to associate a search session with a list of keywords, against which matching will be performed. A webpage will be saved locally only if its page content contains one or more of the user-specified keywords.

4.4. Monitoring Mechanism

BTM provides a convenient mechanism for law enforcement agencies to set up their monitoring operation with established schedules. With the auto-scheduling module, the process (searching torrent files followed by torrent analysis) reruns itself in a periodic manner without users' intervention. In addition, BTM can be configured to search only updated (with new posts since the last operation) topics, and to analyse torrent files, which have not been processed previously. The auto-scheduling and history-shrewd settings make 24 x 7 monitoring possible, efficient and effective.

5. Torrent analyzer

On the completion of torrent file search, torrent files will be downloaded and archived locally in the BTM System. The relay baton is then passed on to Torrent Analyzer, which will communicate with trackers and peers using the BT protocols to gather information in the swarm. With the torrent files, Torrent Analyzer connects to the trackers and retrieves information about the lists of peers currently participating in the download. These peers are then contacted for details about their download status. Responses from trackers and status of peers will be analyzed together to generate useful information and statistics, which will be conveyed to BTM users in the analysis report.

Torrent Analyzer can be configured to dig out specific information about torrent files and the connected peers. To cope with this, a rule system, to be applied to the torrent files and peers, is developed and has been incorporated into the BTM System. An alert will be prompted (in error message dialogs) or sent to pre- specified email addresses when a certain rule is satisfied. An overview of the Torrent Analyzer is given in Figure 5.

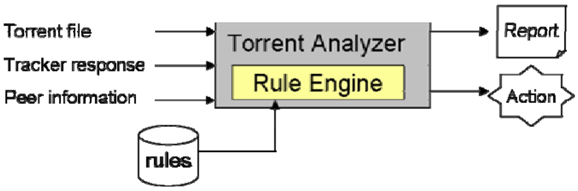


Figure 5. An overview of the Torrent Analyzer

5.1. Rule System Employed by BTM

In the rule system employed by BTM, attributes are the fundamental units used for matching. Attributes are basically classified into two categories: torrent-attributes and peer-attributes. They are associated with the torrent as a whole and the peers respectively. Definitions and the supported operators on these attributes are summarized in Table 1 (torrent attributes) and Table 2 (peer attributes).

Table 1. Attributes defined for a torrent file *T*

Attribute	Definition	Operators
Created By	Client program by which *T* is created	'is', 'isn't', 'contains', 'doesn't contain'
Torrent Name	Name of the files to be distributed by *T*	
Creation Date	Date on which *T* is created	'is', 'isn't', 'before', 'after'
Seeder Country	Country in which connected seeders are located	'is', 'isn't'
Non-Seeder Country	Country in which connected non-seeders are located	
Number of Seeders	Number of connected seeders associated with *T*	'=', '!=', '<', '>'
Percentage of non-seeder obtain	Percentage of file of which non-seeders have obtained.	

Table 2. Attributes defined for a peer *P*

Attribute	Definition	Operators
BT Client	Name of the BT client used by *P*	'is', 'isn't', 'contains', 'doesn't contain'
IP Address	IP Address of *P*	
ISP	ISP of *P*	
Country	Country in which *P* is located	'is', 'isn't',
Percentage	The percentage of file(s) available at P	'=', '!=', '<', '>'

A rule may consist of one or more conditions, each of which is made up of an attribute, an operator, and a value. Peer-conditions and torrent-conditions refer to the conditions describing peer attributes and torrent attributes respectively. Peer-rules and torrent-rules refer to the rules involving peer-conditions and torrent-conditions. Within the same rule, conditions must be of the same type and are processed conjunctively.

When multiple rules are defined in a session, they are applied separately to each connected peer or to the whole torrent, according to the attributes involved. Peer-rules are evaluated on each peer once it is connected while torrent-rules are checked after all peers' statuses have been gathered.

5.2. Rule Execution Algorithm

The rule engine is similar to traditional rule-based expert system discussed in the artificial intelligent literature. The engine consists of three main components: a working memory, a rule interpreter and a collection of rules. In BTM, the working memory (WM) stores the information associated with the peer attributes and the torrent attributes. The collection of rules (*Rs*) is the list of rules described in A. The rule interpreter uses the following algorithm for execution:

```
For each rule $R_i$ in $Rs$
    For each condition $C_{ij}$ in $R_i$
        If $C_{ij}$ matches with an element in WM
        Then continue with next condition
        Else continue with next rule
    If all conditions in $R_i$ match successfully
    Then execute the action $A_i$ in $R_i$
```

5.3. Illustration of Rule Evaluation

The using of rule-based approach in BTM allows different types of users to have the flexibility to program BTM with different behaviour. For example, law enforcement officers may want to identify seeders of piracy works in their jurisdiction whereas academic researchers may want to collect statistics on piracy within the Asia region.

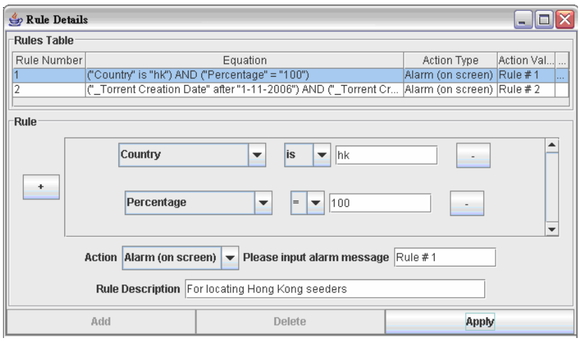
The following rule, RULE #1, is used to identify seeders in the country 'HK', which may be used by law enforcement officers to trace potential publisher of piracy work.

- RULE # 1
  Condition 1: Country *is* 'HK'
  (The peer is located in HK)
  Condition 2: Percentage *is* '100' %
  (The peer has obtained 100% of the files→ a seeder)

The second example, RULE #2, attempts to find all distinct torrents that were created in November 2006 and with seeder from 'HK'. The result provides information on the possible extent of distribution of piracy work in November of 2006. If we define one rule for each month of the year, we are able to study the trend of piracy within those selected web sites in the year.

- RULE # 2
  Condition 3: Creation Date *after* '1-11-2006'
  (The torrent file was created after 1-11-2006)
  Condition 4: Creation Date *before* '30-11-2006'
  (The torrent file was created before 30-11-2006)
  Condition 5: Seeder Country *is* 'HK'
  (The seeder(s) of this torrent is located in Hong Kong)

The following example illustrates how BTM evaluates the rules. Before the start of the operation, the above two rules are defined as illustrated by the sample screen dump of the BTM system in Figure 6.



Rule Details

Rules Table

Rule Number	Equation	Action Type	Action Val...	...
1	("Country" is "hk") AND ("Percentage" = "100")	Alarm (on screen)	Rule # 1	...
2	("_Torrent Creation Date" after "1-11-2006") AND ("_Torrent Cr...	Alarm (on screen)	Rule # 2	

Rule

Country is hk

Percentage = 100

Action Alarm (on screen) Please input alarm message Rule # 1

Rule Description For locating Hong Kong seeders

Figure 6. A sample screen dump of the BTM system – Rule details

Assume 10 peers are identified based on the tracker's response. BTM begins its analysis by contacting the peers for their download statuses one after another. Rule # 1 will be applied on each peer immediately once it is connected. Necessary information will be gathered from the connected peers and will be used for the evaluation of Rule # 2 later on. Table III shows the relevant information gathered after contacting each of the peers on the peer-list.

Table 3. Information Gathered from Connected Peers

Peer ID	001	002	003	004	005	006	007	008
Country	cn	hk	hk	cn	ca	ca	hk	hk
Percentage	55	30	100	0	40	80	45	45

In this example, Peer 003, a seeder *(Condition 2)* located in Hong Kong *(Condition 1),* will trigger an alert when Rule # 1 is applied on it. From the torrent file, we found that it was created on November 13, 2006. After all these peers have been evaluated, Rule # 2 will be applied and this torrent will trigger the corresponding alert as the creation date is after November 1, 2006 *(Condition 3),* and before

November 30, 2006 *(Condition 4)* , plus the seeder is located in Hong Kong *(Condition 5).*

5.4. Accuracy

Data obtained from a peer is just a snapshot. The situation in a swarm changes continuously as peers come and go. In other words, even if an alarm has been triggered at a certain moment, the peer may not match the triggered rule a while later. For this reason, a verifying timestamp is added to every alert. For torrent-rules, the accuracy may be even lower. In particular, the following situations have significant influence on the accuracy:

High bandwidth peers

If a peer has a high bandwidth connection and fast uploaders or seeders are connected, the download speed of the peer can be extremely fast. In this situation, the percentage of the files acquired increases tremendously.

Small file size

The percentage of the files acquired increases rapidly when the file size is small, similar to the situation with high bandwidth.

Super seeding mode

On the start of a file transfer, much time can be wasted because the seeding client may have to send the same file piece to many different peers, while other pieces have not yet been downloaded at all. Therefore, super seeding is designed to help a torrent initiator with limited bandwidth to "pump up" a large torrent, reducing the amount of data it needs to upload in order to spawn new seeds in the torrent. Under super-seeding mode, the seeder masquerades as an ordinary peer with no data, and tells the peers that it received a piece that was never sent. As a result, the existence of seeders may not be established.

Partial peer-list from tracker

When a tracker is contacted for the list of peers, we have no control over the set of peers returned. The number of peers to be contacted by BTM depends on the total available peers on the trackers and the number of peers specified in its request. Consequently, attributes ‘Non-Seeder Country’, ‘Seeder Country’, ‘Number of Seeders’ and ‘Percentage of non-seeder obtain’ describe the swarm composed of the partial peer-list only. To guarantee an adequate performance level, BTM restricts itself to track a maximum of 50 peers each time.

6. Conclusion

We have performed a preliminary evaluation on the system. The target forum (sub-forum) for the testing is http://www.uwants.com (Cantonese pop music resources sharing). Using only about 5 minutes, the Torrent Searcher has checked 124 threads and downloaded 114 torrent files. The Torrent Analyzer has successfully identified over 2500 trackers and 3000 peers within 90 minutes. In this exercise, 126 seeders were found. It shows that the system is effective and can enhance the effectiveness of the law enforcement agencies. We expect a significant crackdown on the increasingly rampant illegal P2P file-sharing activities after our Customs and Exercise Department puts the BTM system into production.

For future work, we remark that the processing time could be improved. In particular, when the forum has many levels of sub-forums, the number of pages to be searched increases drastically. It is desirable to develop algorithms to speed up the searching or prune unpromising links more promptly.

7. References

[1] T. Karagiannis, A. Broido, N. Brownlee, K. Claffy and M. Faloutsos. Is P2P Dying or Just Hiding? In *Globecom,* Dallas, TX, USA, November 2004

[2] J.A. Pouwelse, P. Garbacki, D.H.J. Epema, H.J. Sips. The Bittorrent P2P file-sharing system: Measurements and analysis. In: Castro M, van Renesse R, eds. Peer-to-Peer Systems IV, *4th Int'l Workshop, IPTPS 2005*. LNCS 3640, Ithaca: Springer-Verlag, 2005. 205-216.

[3] A. Parker, Peer-to-Peer in 2005, CacheLogic Research, available at http://www.cachelogic.com/home/pages/studies/ 2005_01.php

[4] BitTorrent Protocol, BitTorrent.org, available at http://www.bit torrent.org/protocol.html

[5] Intellectual Property Department, The Government of Hong Kong Special Administrative Region, Awardness of Protection of Intellectual Property Rights Increases, available at http://www.info. gov.hk/gia/general/200601/19/P200601190149.htm.

[6] Anti-piracy Fact Sheet, Asia-Pacific Region, Motion Picture Association, Available at http://www.mpaa.org/AsiaPacificPiracy FactSheet.pdf

[7] IFPI External Press Pack, April 2004, available at http://www. pladebranchen.nu/publikationer/1000nyepiratsager_IFPI.doc

[8] D. Sinn, Downloaders shrug off court threat, South China Morning Post, January 27, 2006

[9] T. Mennecke, File-Sharing Winners and Losers of 2005, Slyck News, available at http://www.slyck.com/news.php? story=1040

[10] D. E. Eastlake, P.E. Jones, RFC 3174 - US Secure Hash Algorithm 1, available at http://www.faqs.org/rfcs/rfc3174.html