BITPREDATOR: A DISCOVERY ALGORITHM FOR BITTORRENT INITIAL SEEDERS AND PEERS

1ST M.SC. RAYMOND BORGES
West Virginia University
Polytechnic University of PR

2ND DR. ROBERT PATTON
Oak Ridge National Laboratory

3RD DR. HOUSSAIN KETTANI
Polytechnic University of PR

4TH DR. YAHYA MASALMAH
Universidad del Turabo

ABSTRACT
There is a large amount of illegal content being replicated through peer-to-peer (P2P) networks where BitTorrent is dominant; therefore, a framework to profile and police it is needed. The goal of this work is to explore the behavior of initial seeds and highly active peers to develop techniques to correctly identify them. We intend to establish a new methodology and software framework for profiling BitTorrent peers. This involves three steps: crawling torrent indexers for keywords in recently added torrents using Really Simple Syndication protocol (RSS), querying torrent trackers for peer list data and verifying Internet Protocol (IP) addresses from peer lists. We verify IPs using active monitoring methods. Peer behavior is evaluated and modeled using bitfield message responses. We also design a tool to profile worldwide file distribution by mapping IP-to-geolocation and linking to WHOIS server information in Google Earth.
KEY WORDS
BitTorrent, Peer-to-Peer, seeder, leecher, initial seeder, content producer

1 INTRODUCTION

In 2009 law enforcement agencies and PROTECT, the national association to protect children, formed an alliance with Oak Ridge National Laboratory to develop software that would assist in saving children from child predators by automating the detective work needed to gather evidence and locate children in danger. In 2009 some work was done to help track pedophilic material being shared on Peer-to-Peer (P2P) networks, specifically; Gnutella based networks such as Limewire and Morpheus. And although these efforts have been far-reaching there are many P2P networks in existence and more appear over time.

The ability to monitor BitTorrent and identify initial content uploaders, seeders or leechers will provide law enforcement agencies a select group of Internet Addresses they further investigate to choose the most valuable targets from. As of now their ability is limited to less used P2P networks such as Gnutella. Even the music and movie industry's ability to monitor BitTorrent is limited to parts of the protocol. They currently track parts of the protocol but it's important to be able to validate results since incorrect addresses can lead to wasted resources or convicting a person of a crime they did not commit.

According to the Internet Commerce Security Laboratory 89% of all BitTorrent transfers are illegal content [1]. This illegal content is not all child pornography (CP) but rather mostly copyright infringing (CI) material. An

investigation into the extent of CP material trafficked on BitTorrent has not been realized by law enforcement agencies, probably because they don't have the tools to do this accurately enough yet.

The software framework and methodology presented here provide tools to perform BitTorrent network content and peer profiling. The methodology provides a starting point for others to design on and improve. It also provides a means to identify patterns and behaviors in the network. It can also provide a tool to help find the global networks formed between content, sharers and downloaders.

2. CLASSICAL BITTORRENT

BitTorrent has a hybrid infrastructure comprised of client-server architecture and P2P architecture shown in Fig.1. Metadata files (called torrents) which ultimately point to the content are obtained through the client-server architecture. Although recently there have been updates to the way they are shared, most are still shared this way today. The protocol relies on centralized servers known as "trackers" to inform peers about each other. It also heavily uses Torrent indexing websites to hold lists of all available files, although torrent files can be shared in many ways from email to flash drives. The protocol follows a logical sequence where the user first locates the torrent metadata file and gets swarm info (peer lists) through the client-server architecture then the user's client application contacts a second web service called the torrent tracker. Finally, the client sorts through the response and contacts peers. This process can be seen in Fig.2.

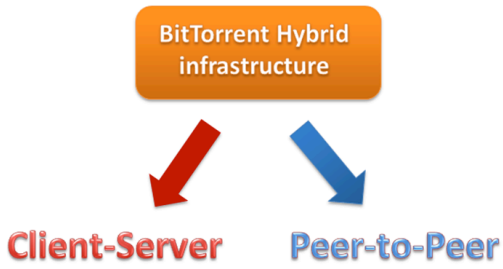

Fig.1 BitTorrent Infrastructure

The infrastructure seen in Fig.2 is composed of 6 elements: three (3) actors, the content to be replicated (this can be more than one file at a time), the BitTorrent client program and a metadata file called the torrent file; the 3 actors are: torrent indexers, torrent trackers and the torrent client user. Torrent indexers are websites that list the metadata torrent files, torrent trackers register peers who are requesting to participate in the swarm (download/upload process) and respond to peers requests for other peers, and users or peers are the members participating in the swarm. Each torrent file, which normally consists of the torrent filename with a .torrent extension may be listed on various torrent indexers and may contains various trackers to obtain multiple peers lists.

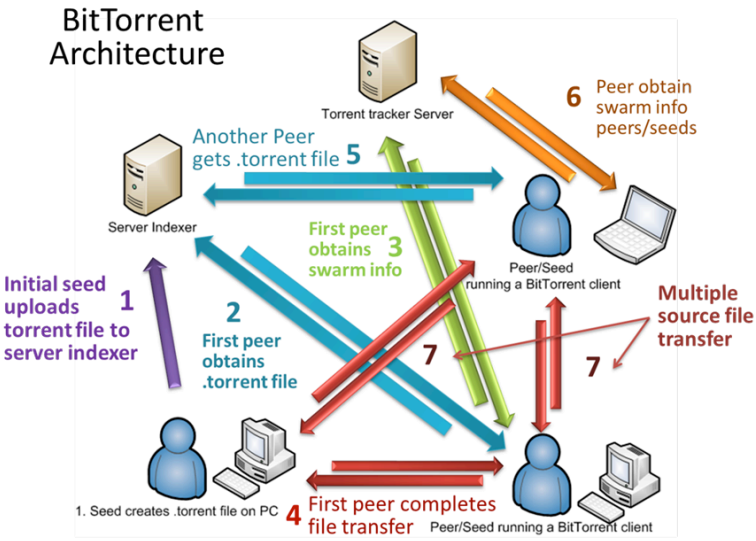

Figure 2. Classical BitTorrent P2P Protocol Sequence

To join a torrent, a peer announces itself to the tracker(s) listed in the .torrent file. This is done by using the Secure Hash Algorithm (SHA-1) hash of the complete info dictionary value which is a concatenation of piece hashes for the entire file. The tracker records the peer's presence and replies to it with a partial list of the peers currently downloading the torrent. The peer then connects to these peers using the BitTorrent peer wire protocol and begins exchanging pieces with them.

3. BITTORRENT CONTENT PROFILING

In previous research it has been found that some methods of monitoring BitTorrent can result in false positives [2].False positives occur when IPs listed in tracker lists are purposely or accidentally inserted. Some torrent trackers have adopted this technique to confound monitors (entities that monitor BitTorrent traffic for copyright infringement). Copyright holders police P2P networks by monitoring P2P objects and sharing behavior; collecting evidence of infringement, and then issuing to an infringing user a so-called Digital Millennium Copyright Act (DMCA) takedown notice. Techniques used by these entities are not suited for specific content distribution monitoring because they tend to use naïve techniques which result in high error rates, some caused by intentional fake IPs being inserted and others due to accidental or simply outdated IPs in peer lists [2].

Some researchers failed to accurately track content because of the tracker server having blacklisted their IP's. This occurs after aggressively monitoring too many torrents at once. They have found that this occurs after the IP used to track the torrents appears in about 100 peer lists [3]. Research suggests that this crawler unsubscribe from each torrent at the tracker after obtaining peer lists in order to prevent being blacklisted.

4. DISCOVERY ALGORITHM

The methodology can be divided into three steps as shown in Fig.3.

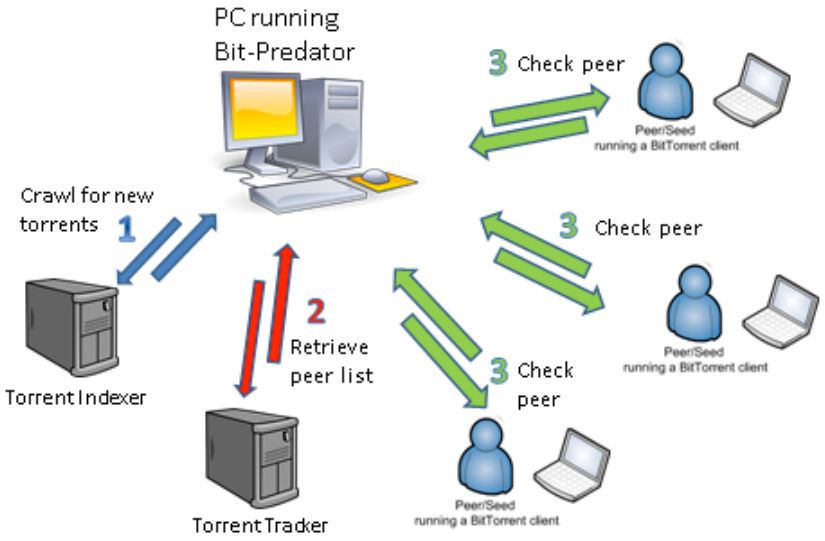

Fig.3. Methodology Overview for BitPredator

.The first step involves collecting the links to the metadata files (torrents). This is a difficult task since torrent files are distributed far and wide over hundreds of torrent indexing websites. This part of the methodology runs on client-server architecture. The software agent developed has to be able to parse Hyper Text Markup Language for torrent filenames and links.

There is also a second technique to acquire torrent links and filenames that uses Extensible Markup Language (XML) and Really Simple Syndication (RSS) feeds which can be parsed. This is the main method used for this research and proved very useful. Either method that is chosen must be set on a time loop to be able to capture the torrent links fast enough to get high value good torrent links

The second step of the first section is to download the metadata file to random access memory (RAM) to parse and determine basic information for the file such as the info-hash and the tracker URL's responsible for tracking this file. The info-hash is the SHA-1 hash of the info section inside of the metadata file. Sometimes the torrent file does not offer all of the information. The protocol for what to list in the metadata file is not followed and variations can be encountered.

The second section in Fig.3 involves contacting the tracker URLs listed in the torrent file. This is accomplished by sending an announce request through HTTP. Contacting each tracker that is reported to track that specific content we can get the swarm info (peer lists of IP-addresses). Once all the trackers are contacted we filter out duplicates and fakes. When that is complete we have a list of suspect IP-addresses where some may belong to real users, others to http

proxies or socks proxies, some random IPs purposely mixed in by trackers to provide plausible deniability to its subscribers and others which may belong to virtual private networks (VPNs) or Tor anonymity network nodes.

The third section of Fig.3 is peer validation. That is, attempting a TCP connection to each of the pre-filtered peers. This is a necessary step to validate that the IPs are actively participating in the torrent share. This is also necessary to be able to discern the type of peer, seeder, leecher or initial seeder.

Once we obtain IP information for the swarm and verify its validity and connection status we need a way to sort this information by location. From freely available a geographic information system (GIS) databases we produce Keyhole Markup Language (KML) files for Google Earth and Google Maps.

5. RESULTS

For situations that follow classical BitTorrent we are 100% sure that we can identify the initial. If the initial seeder deviates from normal behavior by perhaps say creating a torrent file of something that already exists or creating a torrent and then giving someone else the job of seeding the file then the technique will fail to capture the person originally responsible; though it may capture the IP-address for the person who seeded it, he may not be the torrent creator. To add to the confusion, the initial seeder or the creator may not be the content producer at all. Say for example the person is uploading a file that his friend created and gave to him.

6. SUMMARY

We have presented a peer discovery algorithm to monitor BitTorrent. By applying the techniques shown we can determine initial seeders, seeders and leechers in various circumstances. Although fake IP-addresses are injected into tracker swarm lists, protocol modifications have been made to how the protocol is used by some clients to make peers more anonymous and even with active monitoring techniques it is possible for the wrong person to be blamed. We offer no guarantee past obtaining the ISPs service set IP as to who could actually be the person responsible for producing the content.

We have developed this methodology to run on the classical BitTorrent protocol. As of today the modifications made to the protocol have not all been taken into account and we believe many exploits still remain to be found

REFERENCES

[1] Layton R. and Watters P., Investigation into the extent of infringing content on BitTorrent networks, Internet Commerce Security Laboratory, 2010.

[2] Bauer K., McCoy D., Grunwald D., and Sicker D., BITSTALKER: Accurately and Efficiently Monitoring BitTorrent Traffic, University of Colorado, Boulder, CO, USA, 978-1-4244-5280-4, IEEE, 2009.

[3] Piatek M., Kohno T. and Krishnamurthy A., Challenges and Directions for Monitoring P2P File Sharing Networks - or- Why My Printer Received a DMCA Takedown Notice, UW-CSE, 2008.