Peer to Peer Torrenting: How Safe is it?

Omar Loudghiri
under supervision of Pr. Mark Allman
Case Western Reserve University
Cleveland, OH
oxl51@case.edu

Abstract—**Peer to Peer torrenting is a way for users to share files directly from their devices. The file is split and then stored in many locations across the hosts (Seeders) and is sent in small partitions by any available peer. This opens the possibility for malicious seeders to alter the files they are sharing in order to hide malware within the portions of the file they are sending to clients downloading the content (Leechers) [8]. This paper explores whether this is a sizeable risk by performing measurments on the reputation of the Seeders through their IP. We also analyze whether there is any recongizable malware in the files that are received.**

Index Terms—**Peer to Peer, BitTorrent**

I. Introduction and Motivation

Peer to peer (P2P) networking has been greatly democratized in recent years thanks to media sharing websites that rely on the decentralization of the P2P network to avoid the cost and management issues that come with hosting a database. The fundamental concept behind a P2P configuration is that it does not have a central governing node that regulates membership to the network. However, it was hard to achieve such level of autonomy in the case that an entry node could not be found. A Hybrid system was therefore created having trackers that keep an account of the existing alive peers that can share a specific file. Trackers keeping track of information about the file being shared allows them a level of control over the integrity of the file. Tracker based torrenting keeps a hash value of each small part of the file and provides them to the torrent client to check that the right file was received by comparing to the hashed value of the file received. However, some claim that it is possible to alter the torrent segments while keeping the same hashed value. [5]. For the purposes of this paper, torrenting will refer to downloading files using a P2P network.

It is not fully understood whether there is a believable threat behind torrenting services and whether the fact that the file can be changed by unknown malicious peers means that the practice is prevalent and poses a danger to the people that download files using P2P [6].

This paper will offer a preliminary look at whether there is questionable doubt behind the idea that torrenting is unsafe by choosing an array of torrents that contain both freely copyrighted data and unduly licensed copyrighted data from sources ranging from the web archive to the pirate bay. The list of seeders and the data downloaded will be subjected to both passive and active measurements to determine whether there is reason to believe the file transfer could be a threat.

The most used torrent clients available (BitTorrent, utorrent, etc.) all have a checksum function that verifies the integrity of the file that is being downloaded, each chunk of data that is sent by each Seeder has a hash value associated with it that then gets authenticated against the tracker's hash value and flags it if there is any discrepancy. The seeder would then be flagged and even banned from using that tracker again. This means that tracker based torrenting has an added layer of security that would prevent malicious seeders from poisoning their torrents. However, up until 2020 these checksums were based on SHA-1 [4]. While SHA-1 does offer a substantial level of protection, it has been demonstrated that some SHA-1 collisions can be exploited and trick the checksum and end up sending files that have been maliciously tampered with. BitTorrent is now using SHA-2 (256 bits) but not all trackers have been updated yet according to the BitTorrent blog.

There are other potential security risks that can arise from being in a P2P network with malicious actors since your IP will be shared with them and they have access to information such as the port you have open for the P2P transfer. These security issues can be explored further in future work, however, it is beneficial to know the amount and percentage of peers deemed suspicious. The suspicion factor of the peers will be evaluated based on many databases that evaluate IPs based on their involvement in known malicious activity. The specific factors will be discussed further in the Data gathering section.

The Hypothesis of this paper is that torrenting from illegal sources poses a higher risk than torrenting from legal sources. However, the expectation is that they both expose the user to non negligible risk. [10]

II. Background

A. BitTorrent

BitTorrent maintains file integrity during sharing by dividing the files into segments and keeping track of the hashed value that corresponds to that segment. Each segment is owned by a seeder who can distribute it based on their preferences and network limitation. Whenever a full segment is received, the torrent client makes sure the hash values are the same to ensure that the segment was received properly. When the hash values do not match, the received segment is discarded and another peer is prompted for that same segment by the torrent client. [4] A segment itself is also split into smaller chunks to facilitate the fragmentation of the data as can be seen in Figure 1.

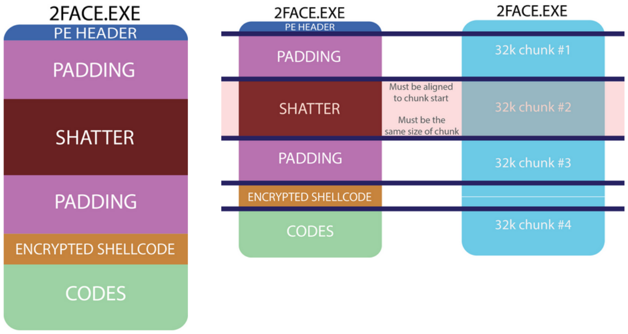

Fig. 1: SHA-1 Collision Expoit. Source [1]

1) BitTorrent Hashing Security flaw: The main security flaw that can be exploited when a file is shared in a P2P network using the BitTorrent SHA-1 security has been proven to work in this GitHub repository https://github.com/skelsec/BitErrant where an encrypted shellcode can be maliciously inserted into a segment as shown above in Figure 1. That modified segment will have the same SHA-1 value if a certain byte arrangement is respected.

This means that it is possible to circumvent the built-in security measure and send malicious code. However, this code is easily picked up by clamAV when the file is scanned, as has proven a test compilation of this code with a file downloaded through torrenting as an input. This means that if such methods were utilized to tamper with a file, an additional layer of security is necessary to detect malicious attempts. In this paper, clamAV will be run on the files downloaded to check for the introduction of malicious code using this technique.

2) Exposure to malicious IPs: While the BitTorrent protocol monitors the integrity of the shared files, it does not monitor the IPs that join a P2P network. Any machine that has access to the tracker can become a node and have the power to seed the file being downloaded. This uncovers the second threat possibility that this paper will be exploring. We will be running all addresses that were part of the networks we joined to download files through databases that keep track of malicious activity from IPs throughout the internet.

III. Data Gathering

In this study, we analyzed 32 torrented files from the sources listed in Table 1 and downloaded them using qBittorrent, a torrent client that supports developer extensions, allowing us to collect the IP addresses of the seeders for each downloaded file.

The list is collected after the end of the download and captures all the peers that have established an incoming connection to our machine and ranks them in descending order by amount of data shared with us (the IP that sent the most data first). It also includes the port that we were connected to on the destination device, however that data is cropped from the list because the IP addresses were needed alone for the databases checks.

Source	# of Torrents	Legality	# IPs
The Pirate Bay	12	Illegal	758
The Internet Archive	12	Legal	521
Libgen.ru	4	Illegal	254
Ubuntu / Kali	4	Legal	309
# of IPs	**Legal**: 830	**Illegal**: 1012	**Total**: 1842

Table 1: Torrent sources

Below, we will explain the criteria the suspicion characterization databases use to list IPs in their database. Five databases were chosen to have comparison points between the data returned and to increase confidence in the repeated cases of malicious indication and help take outlier into less consideration when that is the case.

A. IP Reputation Lists

The databases that consolidated information about the peer IP addresses were taken from a Threat Intelligence resource list [7] which gathers up-to-date evidence-based resources on IP addresses. While the absolute legitimacy of each database could be debated, we chose to use results from many sources to attempt to reach a higher confidence level.

The first four databases provide their own API which was used for the data collection of this paper. Trial accounts and free accounts were sufficient to run all 1842 IPs through each database at least twice over a three week period. The fifth database is a list hosted on GitHub which can then be downloaded and cross checked with the collected IPs using a script.

1) AbuseIPDB: According to their self description, AbuseIPDB is a project that works to prevent hackers, spammers, and abusive activity on the internet. It provides a central blacklist for webmasters, system administrators, and other interested parties to report and find IP addresses associated with malicious online activity. [7]. They are sponsored by two companies:

i Security Trails which claims to have the largest repository of DNS and WHOIS records. It also provides a forensic tool that investigates malicious IP addresses.

ii IP2Location which gives its GeoIP database to AbuseIPDB to improve information on malicious IPs.

While these two companies provide a solid ground for this database's data, they also allow registered sysadmins to report malicious IPs to the database. The ranking system is a malicious score over 100 (0 is not malicious/not in database and 100 is completely malicious). Each sysadmin can report an IP once to increase their maliciousness score by 5, the reports reset after 60 days if no additional other reports are filed. Common users can also report an IP once every 60 days to increase its score by 1 and that point also resets every 60 days. When an IP reaches a certain threshold that is not specified, it is further investigated and if deemed malicious the score takes 365 days to reset. [1]

2) ThreatJammer: The Threat Jammer database gives us access to various other threat intelligence databases. It includes data form both open source and closed source threat intelligence databses. The list of all the databases it uses can be found at https://threatjammer.com/osint-lists. Notable ones are CINS Army, Blocklist Project, and AHA (Anti Hacker Alliance) ... We made sure to exclude AbuseIPDB from this data point in our queries to avoid intersecting data. All the databases in the IP threat reporting list claim to update every hour.

3) Focsec Database: Focsec.com offers an API that can identify VPNs, proxies, bots, and TOR requests. By providing real-time updated data, this tool identifies potentially suspicious logins, malicious activities, and abusive or illegal behavior. According to their user forum, they use data from Abuse.ch's SSL BlackList, which keeps track of malicious SSL activity and keeps track of IPs involved with that that activity. Focsec claims to get updated data from various sources including their own data probing and risk assessment.

4) Project HoneyPot: Project Honey Pot is a user run project that runs script on websites to track IPs that are associated with spam activity, illegal data harvesting, and malicious usage of that data. This is particularly relevant to torrenting since the IPs that are involved in malicious harvesting of data can also be present in P2P networks and harvest data from other nodes in the network.

5) IPsum List: IPsum is a threat intelligence list that collects and combines over 30 public repositories of potentially harmful IP addresses. The lists are updated every day and the resulting collection is stored in their GitHub repository. It shows each IP address and how many times it appears on the several blacklists. The list contained 228,409 IPs at the time of first collection (April 12th, 2023) and then redownloaded on May 2nd, 2023 and contained 228,355 IP addresses. The collected IPs were then crosschecked with the peer IP list.

B. ClamAV and bitErrant

ClamAV is a command line based anti-malware tool that can either run passively on a device or actively scan input files. Version 1.1.0 was used for this project and the malware reference database was run before each scan. While ClamAV has a reported accuracy of 75.45 %, it compares to other paid services since the highest accuracy found in the comparison done by ShadowServer was 80.28% [3] Cisco is now the owner of ClamAV and in turn oversees keeping its database current with new emerging malware. The virus signature database is reportedly updated every four hours.

For the purposes of this paper, we verify that ClamAV can identify files that were tampered with using the bitErrant [9] code. The BitErrant SHA-1 exploit would be one of the few identified ways to circumvent BitTorrent's layer of security. We attempt to modify 5 files that were downloaded using qBitTorent which performs SHA-1 scan. While we only added noise data and no executables using the script, .mkv and .pdf files were now different and couldn't be opened by their normal software. The script cited above performs a SHA-1 test and claims that the input and output have the same SHA-1.

All the compilation instructions are present in the script's github repository. We are unsure whether this would in fact circumvent a torrent client's hash check since there are clear changes to the output file. The effectiveness of the bitErrant script is a topic that could be explored in further works, and this was only a proof-of-concept test.

Running clamAV on the 5 input files that previously did not return any flags. However, when running clamAV on the output files after using the bitErrant script, the injected noise is detected as:

```
/Donwloads/Parallels-Desktop-Business-Edition-
18.0.1+Crack(macOS): HEUR:Heuristics.Corrupted
```

This means that ClamAV's detection algorithm has detected that the file was corrupted based on a heuristic specially designed to detect corrupted files.

C. IP Geo-location

The IPdata API is used to determine the location of the IP addresses that were flagged as malicious by the above databases. This API is used by many well-known companies and research institutes for data collection purposes [2]. It also provides with a free trial that allowed us to collect locations for all our malicious IPs.

D. Focsec IP classifications

In addition to providing data about the risk index of a specific IP the focsec API also provides data about whether a specific IP is:

1. A proxy. 2. Used as a VPN. 3. A Tor node. 4. A data center. 5. Part of a bot net.

This data will be used to to check if an IP:

- Is identified as a proxy or VPN, the actual malicious actor may remain unknown as they could be using a different IP address that has not been flagged.
- Actively involved in potentially illegal activity on the darknet.
- Determine whether the malicious actors are in fact data centers that not only engage in malicious activity but host more user data and therefore have access to sensitive information.
- Is a compromised device that can be involved in malicious activity without the knowledge of the owner.

IV. Data Analysis And Results

After our data collection and according to Table 2, we can see that we encounter 11.6 % more peers flagged as malicious when downloading illegally copyrighted files from sources that are considered illegal and riskier. However, while downloading legal content there is always a non-zero percentage of peers that is flagged as malicious. The overall presence of peers that

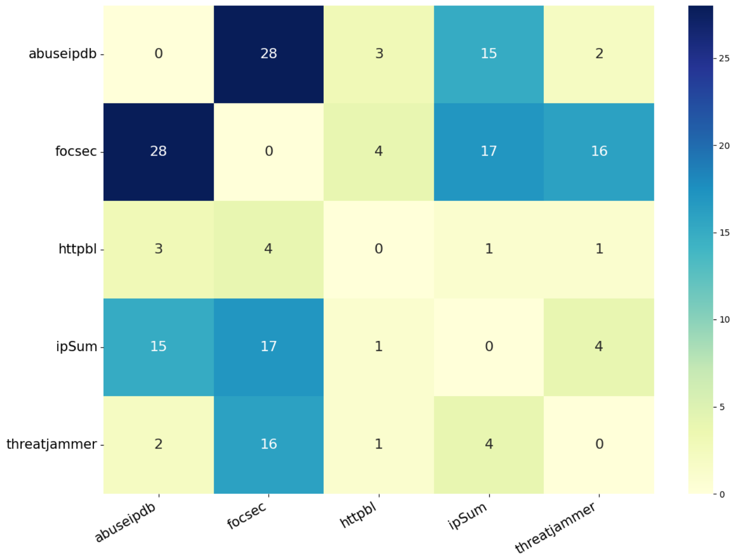

Fig. 2: Counts of malicious IPs intersecting in each database

Folder	Legal	Illegal	Count
Abuseipdb	5	35	40
Focsec	13	38	51
httpbl (Honey Pot)	4	19	23
ipSum	2	35	37
Threat Jammer	6	27	33
Total	30	154	184
% of collected IPs	3.6%	15.2%	9.9%

Table 2: Origin of Suspicious IPs

were flagged as malicious by at least one database is 9.9% across both legal and illegal kinds of torrenting.

Looking at Table 3, the average threat score of a malicious IP from both the abuseIPDB and ThreatJammer databases is 19.72 points higher for illegal downloads compared to legal downloads and the maximum value is 41 points higher for illegal sources.

These results preliminarily confirms our hypothesis that torrenting exposes users to a security risk because of how easy it is for malicious users to access. It also means that it is safer to torrent legal material. Malicious users usually target illegal sources because of various reasons that could range from the lack of legal resources the user engaging in illegal activity might have once affected, to lack of moderation of the trackers. Many open-source software communities with engaged moderators and contributors offer torrenting as a download option which is therefore well monitored by these contributors. While it can be seen that piracy communities across the internet (reddit, 4chan, etc) also benefit from great engagement, the amount of pirated content is magnitudes larger than the legally torrentable content on the internet. To verify the threat reports from certain database, we need to cross check that various unrelated sources agree on the level of danger a certain P2P user poses.

A. Database Intersections

Statistic	Legal	Illegal
Average	16.25	35.97
Median	17.50	33.00
Std. Dev.	14.47	26.05
Min	1.00	1.00
Max	57.00	100.00

Table 3: Legal and Illegal Downloads Threat Scores

The databases introduced in section 3 all have different primary and secondary sources they gather their information from. Figure 2 shows a heatmap of how often any two databases intersect, while Table 4 shows the amount of times two or more databases agree with each other. We can observe that at least two databases agree for 40.2% of the malicious IPs. The heat map shows that the pairs of databases that agree most, in descending order, are: IPsum and Focsec (17), Focsec and Threat Jammer(16) and , IPsum and AbuseIPDB (15). All other databases agree on less than 10 IPs each.

The average threat score of the IP addresses that appear in 2 or more databases is 30.55% which is less than a standard derivation away from both the legal and the illegal average . This means that when we get data from agreeing databases, it does not impact the threat score in a statistically significant way.

There also 2 IPs for which every single databases agree:

```
45.131.195.39 [threat score: 19]
154.47.25.171 [threat score: 49]
```

While all 5 databases agree that these two IPs constitute a threat, the fact that they are 81 and 51 points away from the highest threat score encountered suggests that the threat scores alone might not indicate as much threat as their presence across different reporting services.

The fact that some databases agree does increase our confidence in the fact that those IPs are more suspicious. The more databases agree, the more confident we are that an IP address is associated with a malicious actor. This means that the report that landed that IP on any of the lists was not just an isolated incident and is based on separated reports of suspicious activity. Constant presence across the databases might indicate a more significant threat level than a threat score alone.

Repeat Count	Occurrences
2	62
3	10
4	10
5	2

Table 4: Occurrences of Cross Referenced IPs

If we consider that the IPs that present a significant risk when in contact with are the ones that are more universally recognized as suspicious.

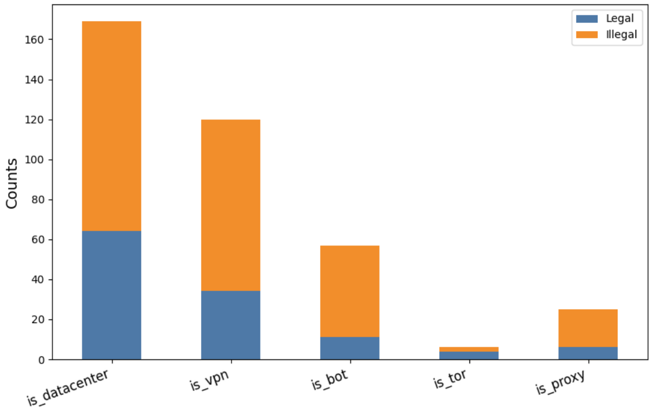

Fig. 3: IP classfication based on Legal/Illegal origin

Out of 184 suspicious IPs, according to this metric and Table 4, only 62 (33.6%) pose a believable threat and even less (10, 10, 2) constitute a more significant threat. Therefore, out of all the IPs encountered only 3.3% have more a significant ground to be believed as threatening since the other isolated reports might be exceptional cases.

B. IP Utilization Flags

In order to gain a greater understanding of the profiles of the IPs we encountered, we will analyze them using the general info Focsec IP database. This database determines a number of attributes (listed in 3.D) about each IP address. We will further analyze the profile of the suspicious IPs in order to understand the utilization and operation of these addresses. Taking a look at Figure 3, we can observe that 8.8% of our seeders are in fact data centers. This significant portions of data centers indicates that some seeders outsource the hosting to specialized facilities which introduces a commercial aspect to otherwise community-based P2P networks. More importantly, according to Table 5, we can see that we were in contact with 31 IPs that were both associated to a data center and flagged for suspicious activity. Individual malicious users already constitute a significant enough threat. Data centers that host data in large scales being associated with malicious activity could mean that users are exposed to greater amounts of compromised and malicious data, that is being massively distributed through these data centers.

IP classification	# of Suspicious IP
Data Center	31
VPN	59
Botnet	30
Tor	6
Proxy	12

Table 5: Amount of Suspicious IP per flag

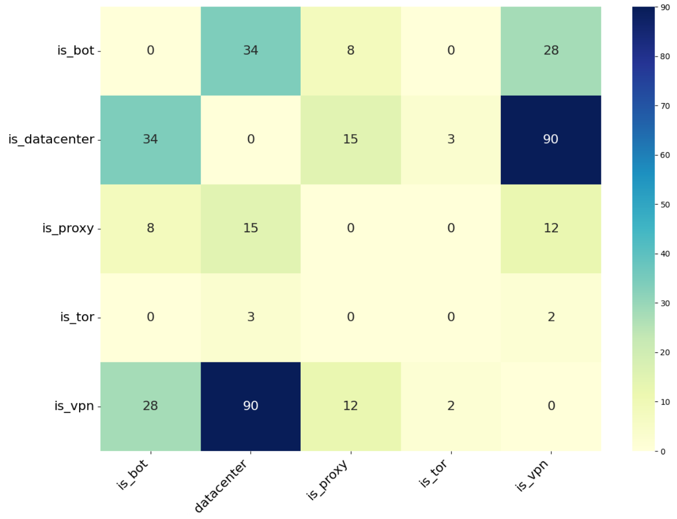

Fig. 4: Intersection of IP classification flags

Moreover, we were able to determine that 32.0% of the suspicious IPs we collected were in fact VPNs. Meaning that it might have been not one but many people with malicious intention that triggered the reports and therefore there is no guarantee such activity might happen again based on that report. On the other hand, it might be that malicious users use already reported VPNs to carry out attacks and inject malware in order to avoid their personal IPs being reported, meaning that the traffic of malicious activity might be even higher for this kind of address.

The other flags are relevant when looking at Figure 4, since we can observe that 51% of the devices part of a botnet are in fact hosted in data centers. This could be a result of either an unknown infection of that device within the data center or an intentional user hosting parts of their botnets in commercial data centers to avoid the malicious traffic going through their physical location. The darkest square of the map indicates that 54% of the data center associated IPs are in fact VPNs. Which means that either the VPN companies outsource the hosting of their remote servers to data centers, or that the data center is directly owned by the VPN companies which might explain the amount suspicious activity reported.

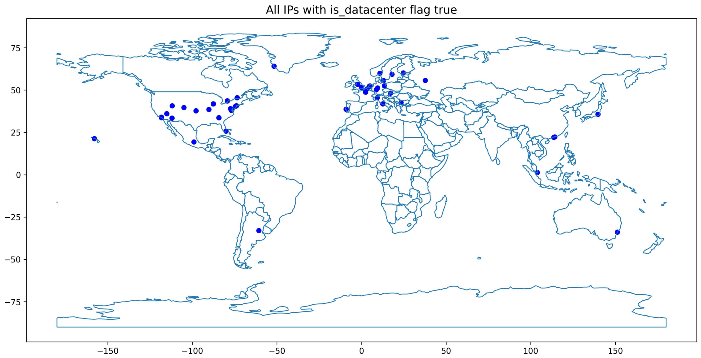

Fig. 5: Geographic location of IPs identified as data centers

There was a suspicion that the IP addresses associated with both VPNs and data centers might have just been a coincidence and located in the same data center hosting files that we happened to select. To verify this, we checked the geographic location of the IPs associated with data centers by plotting coordinate data onto Figure 5. The results showed that the data center IPs were not clustered in a single location. However, it is possible that the suspicious IPs collected were in fact VPN nodes that many people have access to, which according to the behavior of a specific user, keep getting reported. It's possible that these VPN nodes are owned by the same VPN company across the globe.

Therefore, we need to be cautious when interpreting the results, as this could be a factor that requires further investigation. In that sense, otherwise harmless VPN users trying to protect their privacy while torrenting might be falsely flagged as a threat because of their IP address when in fact it was another user that caused the previous report that prompted that IP's threat score to go up. Therefore showing that up to 32% of the suspicious IPs we identified might just normal users using VPNs.

C. ClamAV

While scanning IPs can be useful to determine whether there is a risk our machine might have been exposed to someone with malicious intents. Even if our previous results lead us to think that we were in contact with a non null amount of verified risk. We also conducted an assessment to determine whether the exposure has affected the file's integrity and whether it has been compromised with malware. Using ClamAV to scan all 32 of the files received yielded no results as all the file scan returned results similar to the ones below:

```
-CHECKSUM: OK
  ----------- SCAN SUMMARY -----------
Known viruses: 8664778
Engine version: 1.0.1
Scanned directories: 1
Scanned files: 2
Infected files: 0
Data scanned: 0.00 MB
Data read: 4626.67 MB (ratio 0.00:1)
Time: 11.536 sec (0 m 11 s)
```

Various custom virus signature databases were loaded onto ClamAV to increase the scope of the scan to no avail. The antivirus was tested against known malware and tampered-with data which both returned the expected positive results as shown earlier in the paper.

While these negative results go against our initial expectations, the consistency of the outcome leads us to believe that none of the files we downloaded contained malware. The sample size might have had an impact on this result, however, suspiciously sourced files were intentionally chosen for the purpose of this paper. We attempted to download files with the least amount of confidence and community endorsement and therefore believe that this result might be partially representative of the absence of malware from torrented files in general.

D. Overall Analysis

Following each of the previous data points presented, we shrunk our pool of believably malicious actors' activity. This paper gathered a moderate amount of evidence to challenge the previously believed claim that engaging in torrenting exposes users to significant risk.

V. Ethics

During the course of this research, it is important to address the ethical considerations that emerged throughout the process. Although some files downloaded were considered illegal, it should be clarified that the intentions were purely for the purpose of scrutinizing and identifying potential malware, rather than engaging in or endorsing any illegal activities. Upon completing the analysis, the files were promptly deleted to ensure that no unethical practices were encouraged.

VI. Conclusion

In conclusion, this paper provides a preliminary analysis of the potential risks associated with torrenting files from both legal and illegal sources. Through the analysis of Seeder IP reputations and checking the files received for any recognizable malware, we attempt to determine the level of risk that torrenting presents.

We have observed that while downloading legal content, there is always a non-null percentage of peers that are flagged as malicious. However, when downloading illegal content, we encountered 11.6 % more peers flagged as malicious.

Further analysis of the collected data allowed us to gain insight into the profiles of suspicious IPs. It was determined that data centers and VPNs constituted a significant portion of the suspicious IP addresses. Moreover, we found that a considerable number of data center IPs were associated with VPNs.

Our results preliminarily confirm that while torrenting exposes users to security risks it is not as significant as thought prior to the paper. Users should still exercise caution when engaging in torrenting.

Future research could focus on verifying the effectiveness of various security measures and tools in mitigating the risks associated with torrenting. Additionally, the study could be expanded to investigate other factors contributing to the security risks in P2P networks and the relationships between the different types of suspicious IP addresses such as data harvesting and future interactions with previously contacted malicious IPs.

References

[1] About abuseipdb. *AbuseIPDB*.

[2] Ipdata. *IPdata*.

[3] Virus180-daystats. *Shadowserver*, 2011.

[4] Bittorrent v2. *Libtorrent Official Release Note*, Sep 2020.

[5] Sha1 collision attack can serve backdoored torrents to track down pirates. *BleepingComputer*, 2020.

[6] Anirban Basu, Simon Fleming, James Stanier, Stephen Naicken, Ian Wakeman, and Vijay K. Gurbani. The state of peer-to-peer network simulators. *ACM Comput. Surv.*, 45(4), aug 2013.

[7] Hslatman. Awesome threat intelligence. *GitHub*.

[8] Rodrigo Rodrigues and Peter Druschel. Peer-to-peer systems. *Commun. ACM*, 53(10):72–82, oct 2010.

[9] Skelsec. Biterrant. *GitHub*.

[10] Xue Zheng, Jun He, Xiaojing Gao, and Kebin Chen. Security analysis of peer-to-peer network topology based on gephi platform: P2p network security analysis via gephi. In *Proceedings of the 2021 2nd International Conference on Control, Robotics and Intelligent System*, CCRIS '21, page 176–180, New York, NY, USA, 2021. Association for Computing Machinery.