UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

ADLER HOFF SCHMIDT

# Characterizing Dissemination of Illegal Copies of Content Through BitTorrent Networks

Thesis presented in partial fulfillment of the requirements for the degree of Master of Computer Science

Prof. Dr. Luciano Paschoal Gaspary
Advisor

Porto Alegre, January 2013

**CIP – CATALOGING-IN-PUBLICATION**

Schmidt, Adler Hoff

Characterizing Dissemination of Illegal Copies of Content Through BitTorrent Networks / Adler Hoff Schmidt. – Porto Alegre: PPGC da UFRGS, 2013.

71 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2013. Advisor: Luciano Paschoal Gaspary.

1. P2P Networks. 2. BitTorrent Networks. 3. Data Sharing. 4. Illegal Movie Copies. I. Gaspary, Luciano Paschoal. II. Título.

*“Everybody is a genius. But if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid.”*

— Albert Einstein

# ACKNOWLEDGMENTS

I would like to start thanking my family for the support they have always given me throughout this journey. To my father that has always been an example of character and generosity. To my grandmother, that never tired herself of praying for good things to happen to me. To my brother, who is and always will be my best friend. To my mother, the strongest and bravest women I've ever met and to whom I dedicate this dissertation.

Also, I would like to thank my professor, advisor and friend Luciano Paschoal Gaspary for believing in me even at my worst moments, those that not even I believed any more. His advises and patience with me were essential to overcome the hard times passed throughout these two years. Without him, I most certainly wouldn't be able to complete this journey.

At last, I would like to express my eternal gratitude for my "co-advisors" and friends Rodolfo Antunes and Weverton Cordeiro. Know that your help was vital for the conclusion of this work.

# AGRADECIMENTOS

Gostaria de começar agradecendo à minha família pelo apoio que sempre me deram ao longo dessa jornada. Ao meu pai que sempre foi um exemplo para mim de caráter e generosidade. À minha vó, que jamais se cansou de rezar pedindo o meu bem. Ao meu irmão, que é e sempre será meu melhor amigo. E a minha mãe, a mulher mais forte e valente que já conheci e a quem eu dedico essa dissertação.

Também gostaria de agradecer ao meu professor, orientador e amigo Luciano Paschoal Gaspary por ter acreditado em mim mesmo nos meus piores momentos, aqueles em que nem eu mesmo acreditava mais. Seus conselhos e sua paciência comigo foram essenciais para superar os momentos de dificuldades passados durante esses dois anos. Sem ele, com certeza não teria conseguido completar essa jornada.

Por último, quero expressar a minha eterna gratidão aos meus “co-orientadores” e amigos Rodolfo Antunes e Weverton Cordeiro. Saibam que a ajuda de vocês foi vital para a conclusão desse trabalho.

# TABLE OF CONTENTS

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| BT | BitTorrent |
| CDF | Cumulative Distribution Function |
| DHT | Distributed Hash Table |
| DVD | Digital Versatile Discs |
| IMDB | Internet Movie Database |
| IP | Internet Protocol |
| ISP | Internet Service Provider |
| NAT | Network Address Translation |
| P2P | Peer to Peer |
| URL | Uniform Resource Locator |
| VIP | Very Important Person |

# LIST OF FIGURES

# LIST OF TABLES

# ABSTRACT

BitTorrent (BT) networks are nowadays the most employed method of Peer-to-Peer (P2P) file sharing in the Internet. Recent monitoring reports reveal that content copies being shared are mostly illegal and movies are the most popular media type. Research efforts carried out to understand the dynamics of content production and sharing in BT networks have been unable to provide precise information regarding the dissemination of illegal copies. In this work we perform an extensive experimental study in order to characterize the behavior of producers, publishers, providers and consumers of copyright-infringing files. This study is based on seven months of traces obtained by monitoring swarms sharing movies via one of the most popular BT public communities. Traces were obtained with an extension of a BitTorrent "universe" observation architecture, which allowed the collection of a database with information about more than 55,000 torrents, 1,000 trackers and 1.9 million IPs. Our analysis not only shows that a small group of active users is responsible for the majority of disseminated illegal copies, as it unravels existing relationships among these actors and characterizes consuming patterns respected by users interested in this particular set of contents.

**Keywords:** P2P Networks, BitTorrent Networks, Data Sharing, Illegal Movie Copies.

# RESUMO

Redes BitTorrent (BT) atualmente representam o método Par-a-Par (P2P) de compartilhamento de arquivos pela Internet mais utilizado. Relatórios de monitoramento recentes revelam que as cópias de conteúdo sendo compartilhadas são, em grande maioria, ilegais e que filmes são os tipos de mídia mais populares. Iniciativas de pesquisa que tentaram entender a dinâmica da produção e do compartilhamento de conteúdo em redes BT não conseguiram prover informações precisas acerca da disseminação de cópias ilegais. No presente trabalho realizamos um extenso estudo experimental para caracterizar o comportamento de produtores, publicadores, provedores e consumidores de arquivos violando direitos autorais. O estudo conduzido é baseado em dados coletados durante sete meses de monitoração de enxames compartilhando filmes por meio de uma das comunidades públicas mais populares de BT. Os dados foram obtidos via emprego de uma arquitetura de monitoração do "universo" BitTorrent, o que permitiu popular uma base com informações acerca de mais de 55.000 torrents, 1.000 rastreadores e 1,9 milhões de IPs. Nossa análise não somente mostra que um pequeno grupo de usuários ativos é responsável pela maior parte do compartilhamento de cópias ilegais, como desvenda relacionamentos existentes entre esses atores e caracteriza os padrões de consumo respeitados pelos usuários interessados nesse tipo de conteúdo.

**Palavras-chave:** Redes P2P, Redes BitTorrent, Compartilhamento de Conteúdo, Cópias Ilícitas de Filme.

# 1 INTRODUCTION

The BitTorrent (BT) protocol is currently the most used option for content sharing over the Internet (SCHULZE; MOCHALSKI, 2009). A recent study by Envisional (ENVISIONAL, 2011) shows that illegal copies of copyrighted content can be found in more than two thirds of torrents registered at one of the most popular BT trackers. Such number reinforces the common sense that BitTorrent is extensively used for sharing of copyrighted files. The same study also indicates that over one third of the illegal copies are movies.

Several studies, such as (ZHANG et al., 2010; LE BLOND et al., 2010; CUEVAS et al., 2010), have been carried out in the recent past to characterize content sharing in BT networks. None of these, however, focused on issues specific to the dissemination process of illegal copies. For example, little is known about trends in the behavior of users who obtain access to the original content in order to create digital copies of it or publish these illegal copies in BT networks. Further, it is unclear to which degree users who create digital copies of copyrighted content are the same ones that make the corresponding copies available in BT communities. Finally, to the best of our knowledge, the characterization of infringing data consumption via BT networks has not yet been addressed. No study has yet considered native aspects of BT networks as well as those proceeded from disseminators' organization while performing a thorough investigation.

Protection mechanisms are necessary to effectively mitigate the dissemination, and consequential consumption, of illegal copies of copyrighted content through file sharing mechanisms. Owners of protected content, in turn, would be interested in developing strategies in order to minimize the possibilities that their property will be copied and published through illegal means. Such goals, however, require the development of a body of knowledge related to the processes employed in the creation and dissemination of illegal copies in file sharing communities.

This work presents a set of results of an experimental study conducted in order to characterize the dissemination and consumption of illegal copies of content through file sharing communities. We seek to identify trends in the behavior of users who generate such copies, those that publish them, the ones that initially seed them into the networks and, lastly, those downloading and forwarding these contents (therefore partially responsible for the dissemination). Our study focused on communities that employ the BT protocol because it is responsible for most of the P2P file sharing traffic over the Internet. Our observations were conducted with traces collected through extensions developed for the TorrentU monitoring architecture (MANSILHA et al., 2011). Since movies encompass a large portion of the observed illegal copies, our study will be focused on this type of content. Our results, however, can be generalized for other types, such as music and software. With traces recording seven months of activities, we obtained 58,633 torrents, 1,153 distinct community usernames, 1,098 trackers and more than 1.9 million IP addresses. We also observed the activities of 623 content digitalization groups.

From the distributed systems operations and management point of view, we believe an important contribution of our work is the presentation of fresh and in-depth characterization results – obtained by means of a long term, large-scale monitoring campaign – of the dynamics behind the dissemination of illegal copies of copyrighted content in BT networks. The results are deemed meaningful to Internet and multimedia service providers, to the film industry, as well as to a community of researchers who investigate strategies and mechanisms to promote a more secure usage of swarm-based content sharing systems. Furthermore, although not the main focus of this work, we do revisit and extend an architecture (proposed in the context of our group), which is tailored to perform active, application-layer protocol monitoring.

The remaining of this dissertation is organized as follows. Chapter 2 presents concepts related to the sharing of illegal copies of movies and discusses related work. Chapter 3 explains the monitoring architecture and how it was instantiated. Chapter 4 presents and discusses the results obtained in respect to producers of illegal copies, while Chapter 5 focuses on the analysis of consumers. Finally, Chapter 6 presents conclusions and perspectives of future work.

# 2 BACKGROUND & RELATED WORK

The first part of this section presents some empirical information about the processes adopted by digitalization groups in order to generate illegal copies of movies. Next we discuss other studies focused on the characterization of content distributed through BitTorrent networks.

## 2.1 Background: Illegal Copies of Movies

Digitalization groups are responsible for creating copies of movies through illegal methods (WIKIPEDIA, 2011a). They are composed by one or more members and claim merit for their activity by adding their pseudonym to the created torrent files. Empirical observations of BT communities indicate that expert users do not recognize a torrent file as trustworthy (and avoid using it) if it does not contain the digitalization group identification. The use of a pseudonym in torrents allows digitalization groups to build a reputation, and groups seem to compete with each other in this respect. Thus, this pseudonym is observed by both content producers and consumers. A digitalization group also seeks to preserve its reputation with two methods: (*i*) ensuring that its pseudonym is not present in copies created by other groups; and (*ii*) guaranteeing that the digitalization process result maintains an expected quality level.

The users decision about downloading (or not) a specific copy is also influenced by the type of digitalization process indicated in the content information. During seven months, we observed the publication of new movie-related torrents in popular communities. Our observation leads to the types of digitalization processes summarized in Table 2.1. Each one is identified by: an acronym; a source (*i.e.*, the media that serves as basis for creation of the illegal copy); and the minimum expected time an illegal copy based on such digitalization method can be found after the original premiere of the movie. Processes in Table 2.1 are ordered according to the expected quality of the created copy. Our observations regarding quality of image and sound are essentially empirical, but firmly supported by comments posted in blogs and community sites. The expected release dates are applicable to the communities we monitored, but might be different in other communities (*e.g.*, private ones). It should be noted that the "DVDRip" process may be performed with sources of higher quality, such as Blu-ray discs. Copies employing sources of higher quality than DVD discs, however, are also identified as created with the "DVDRip" process.

Table 2.1: Digitalization processes

| Acronym | Source | Estimated Time |
|---|---|---|
| CAM | Recorded at a movie theater | Aprox. 1 Week |
| TS | Recorded at a movie theater with exclusive audio source | Aprox. 1 Week |
| TC | Directly copied from theaters media | Aprox. 1 Week |
| PPVRip | Content exhibited to hotels clients | Aprox. 8 Weeks |
| SCR | Copy distributed to critics and special users | Unpredictable |
| DVDScr | DVD distributed to special users | Aprox. 8 Weeks |
| R5 | Non-edited DVD, launched only on region 5 | Aprox. 4 Weeks |
| DVDRip * | DVD distributed to general public | Aprox. 10 Weeks |

* Sources with higher quality, such as Blu-ray discs, were also considered DVDRip.

## 2.2 Related Work

Studies related to ours are divided in two classes. We first present a summary of proposed monitoring infrastructures for BT networks. Next, we review studies that focus on the observation of the BT "universe" in order to identify its general characteristics and to model the creation, distribution and consumption of content.

Bauer *et al.* (BAUER et al., 2009) proposed a monitoring infrastructure based on active measurement of BT swarms. The monitoring consists in contacting trackers to obtain IP addresses from peers and then verifying these in order to acknowledge them as valid BT peers. Junemann *et al.* (JUNEMANN et al., 2010) developed a tool to monitor distributed hash tables (DHT) associated with BT swarms. This tool is composed of three modules. The first allows the collection of data from the P2P network such as the number of peers and IP addresses and ports through queries to the DHT overlay. The second module analyzes the data and generates graphs according to predefined metrics. The third and final module generates warnings for situations such as torrents with high number of connected peers. Another monitoring infrastructure, named BTM, is presented by Chow *et al.* (CHOW et al., 2007). It focuses on the detection of piracy through automatic monitoring of BT swarms. The BTM architecture is organized in two modules: one for searching torrent files and the other for the analysis of their contents. The characteristics of the pirated content BTM should look for are defined by the user as a set of rules that are employed during the analysis of the collected data.

Studies that focus on a general characterization of the BitTorrent "universe" include the work of Zhang *et al.* (ZHANG et al., 2010), that analysed five public communities by investigating networks traces generated by directly communicating with trackers and monitoring DHT networks. Authors present, among other results: which are the main BT communities; the participation degree of each torrent publisher; the loads and localization of most used trackers; the geographic distribution of peers; and the most used BitTorrent implementations. Similarly, Zhang *et al.* (ZHANG et al., 2010) present an investigation about "darknets" in BT. These are private communities accessible only through subscription and the possible source of initial distribution of illegal copies. Among the results, authors compare characteristics of swarms promoted by darknets against ones from public communities.

Studies that focus on content dissemination and consumption in BT networks include the work of Blond *et al.* (LE BLOND et al., 2010), which presents an analysis of 103 days of monitoring of three popular BT communities. Its results show a profile of the most active content providers and consumers. Authors were able to identify 70% of providers, list the most popular contents being shared and characterize the most active participants (users present in most swarms). Cuevas *et al.* (CUEVAS et al., 2010) study BT networks socio-economical factors, emphasizing the incentives that drive content providers. Three groups of publishers are identified: those who distribute content due to financial incentives, those who act due to altruistic motivation and those who are responsible for fake content. Based on the analysis of one month of traces, the groups are characterized according to: the ISPs to which they are associated; types of content that are published; incentives for their activity and an estimation of possible monetary incomes.

The aforementioned studies from Zhang *et al.* try to quantify and model BT swarms through creation of "snapshots" of their lifecycle. Their scope, however, did not include an analysis about the patterns and dynamics involving strictly illegal copies of content. Considering the employed monitoring techniques and the results granularity, the studies by Blond *et al.* and Cuevas *et al.* are the ones most similar to the one presented throughout this work. Results presented in these studies, however, do not present the necessary information to allow the identification of trends in the behavior of users that disseminate and consume illegal copies of copyrighted content. Issues that directly influence the comprehension of processes used on the distribution of such copies are left unexplored. It is also important to note that these studies do not seem to consider technical issues such as the filtering out of polluted content, which is characterized by torrents with very short lifetime in the community. Such torrents should be carefully to guarantee that their content will not generate spurious results.

In this section we reviewed the most relevant studies that are related to the one presented throughout this work. There are efforts from the P2P research community in order to create the necessary monitoring tools and proceed with observations to better understand the BT "universe". None of these studies, however, focused on understanding how BT networks are used for the dissemination of illegal copies of copyrighted content. Knowledge about these still unknown dissemination and consumption dynamics can support the development of effective strategies and mechanisms for the protection of copyrighted content. It can also contribute to stimulate the adoption of BT networks for legally enforced commercial activities. We believe that such benefits are a sound reason to justify further investigation of the proposed topic. To the best of our knowledge, this is the first study that focuses on systematically mapping the dissemination and consumption process of illegal copies of content in BT networks. The following chapters present the employed monitoring architecture, its instantiation and the most relevant results that were found.

# 3 MONITORING INFRASTRUCTURE

The resulting volume of files being shared in BitTorrent communities is huge (SCHULZE; MOCHALSKI, 2009). To make our observations possible (about illegal distribution and consumption of copyrighted material), we implemented and instantiated a dedicated monitoring infrastructure. This enabled us to keep track of thousands of swarms. The resulting infrastructure allowed us to collect traces of 58,633 torrents, 1,098 trackers and more than 1.9 million IP addresses in a timespan of seven months (from 05/2011 to 11/2011). The employed monitoring architecture, TorrentU (MANSILHA et al., 2011), and the extensions developed in order to allow the required observations are presented in section 3.1. Section 3.2 presents information about the instantiation of the monitoring and the execution of our experiments.

## 3.1 TorrentU Monitoring Architecture

TorrentU (MANSILHA et al., 2011) is a flexible architecture designed and developed for monitoring BitTorrent networks. As presented in Figure 3.1, the architecture follows the classic manager/agent approach and thus basically contains two elements: an Observer and Telescopes. The observer acts as the manager of the architecture. It is a front-end that allows the operator to configure the system and observe the collected results in real time (and also historic data). Telescopes, in turn, act as agents. They are the components responsible for monitoring the BitTorrent universe and returning results according to requests received from the Observer. Telescopes are further divided in three components named “lenses”, each one responsible for monitoring a different group of elements from the universe: communities, trackers and peers. Such modularization allows existing lenses to be changed and also new ones to be easily incorporated into the architecture without modification of other essential components.

Taking advantage of the flexibility provided by the TorrentU architecture, new functionality (not originally envisaged) was implemented and integrated, resulting in two main extensions. The first, created to allow identification of the first seeders of a swarm, is a **seeker lens** that captures torrents as soon as they are published in a community. The second extension is a **torrent lens** that continuously monitors the community Web page (the ones containing information about the captured torrents) in order to collect: swarm lifetime; number of seeders and leechers; and comments

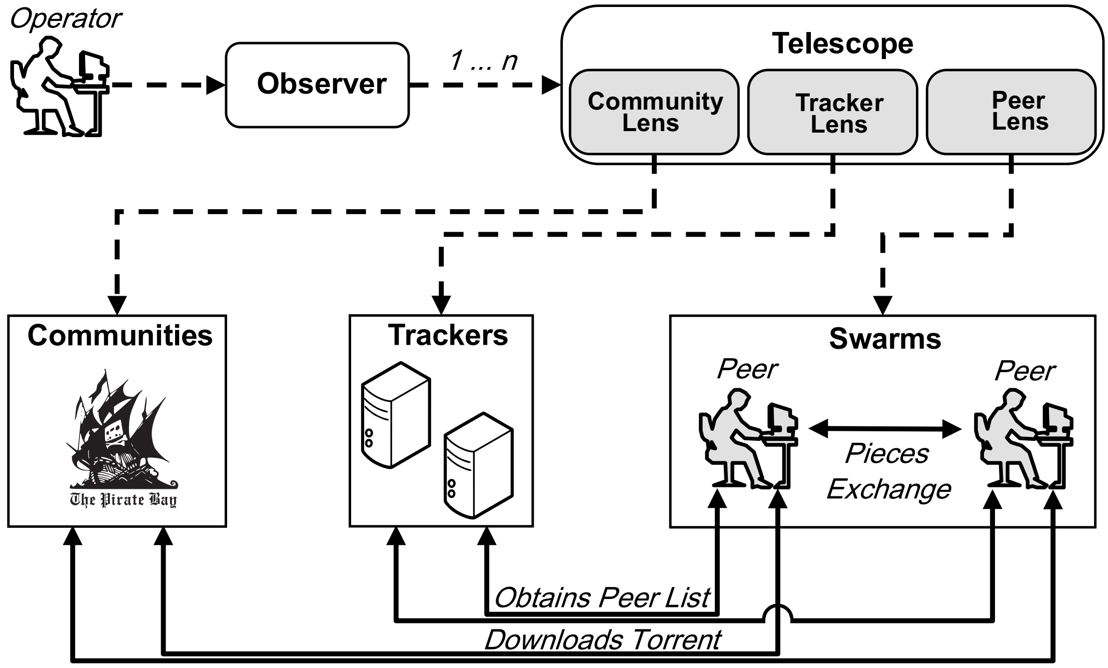



Figure 3.1: TorrentU architecture

posted by users about the content. The goal of the second extension is to create snapshots of the swarm throughout its lifetime.

Algorithm 1 presents a general view of the monitoring process. The first step consists in capturing recently published torrents from communities. For each torrent, a characterization of the tracker is performed and then it is contacted with a request for the peer list of the swarm. If the list is successfully received, a characterization of the first peers participating the swarm is performed. Next, the torrent lens is initialized for the processed torrent. Finally, if the torrent meets minimum requirements, *i.e.*, a flag signalizing the use of peer lens and the successful identification of the first seeders, a set of peer lenses are instantiated. They will continually monitor the swarm so a characterization of its lifetime development can be performed. It should be noted that no content is downloaded during the monitoring: peers are only contacted for acquisition of their bitfields.

There are six parameters that control the execution of the algorithm: *time* determines the duration of the whole monitoring campaign; *attempts* defines the number of connection attempts to trackers; *quantity* represents the size of the peer list requested to a tracker; *threshold* defines the number of peers of a swarm that will be contacted for acquisition of their bitfield; *frequency* defines the time interval between snapshots of a swarm; and *interval* represents the waiting time between each iteration of the algorithm.

It should be noted that this dissertation focus lies in the results of characterizing illegal movie copies dissemination and consumption, and not in the description of the monitoring architecture. The reader interested in more information about the architecture should refer to (MANSILHA et al., 2011).

```
input: time, attempts, quantity, threshold, frequency, interval,
       monitorSwarmGrowth
for i ← 0 to time do
    list[torrent] ← CaptureRecentTorrents();
    for j ← 0 to list[torrent].size() do
        torrent ← list[j];
        DownloadTorrent(torrent);
        ReadFile(torrent);
        CharacterizeTrackers(torrent);
        peerList ← GetPeerList(torrent, attempts, quantity);
        CharacterizePeers(peerList);
        if peerList.size() < threshold then
            ExchangeBitfields(torrent);
            if monitorSwarmGrowth && firstSeeders[].size() > 0 then
                BeginSwarmMonitoring(torrent);
            end
        end
        BeginSnapshotCapture(torrent, frequency);
    end
    Wait (interval);
end
```

**Algorithm 1**: Monitoring process

## 3.2 Architecture Instantiation

The performed monitoring of infringing contents being shared via BT networks was divided on two major phases. Initially, we focused on the aspects concerning publishing and distribution of these contents. This phase lasted a total of seven months worth of monitoring, from 05/2011 to 11/2011. Next, aspects to characterize the consumption of these illegal contents were sought out. This second phase lasted a month and it was on an overlapping period with the prior phase during the month of 11/2011. The decision to tackle this problem with divided phases was taken so that incrementally richer result sets could be examined, processed and presented.

For this monitoring to be possible, Telescopes were deployed in three nodes of the PlanetLab testbed (PLANETLAB, 2011) and in a private server. The Observer component, in turn, was deployed on a single workstation.

Among existing open BitTorrent Communities, we chose PirateBay (PIRATEBAY, 2011) due to its popularity. The Web pages of this community contain only links for torrents that are published through their servers. They also present statistics about users that published a torrent and provide user classification based on his/her reputation in the community.

In order to follow the afore mentioned strategy to an incremental approach, during the first six month a negative value was assigned to the parameter that determined whether to observe swarms lifetime development (*monitorSwarmGrowth*) and, afterwards, was changed to a positive value for the last month. Aside from this parameter, others were identical during both stages and determined as follows:

duration of monitoring (*time*): 7 months; attempts to contact a tracker (*attempts*): 2; number of peers requested from trackers (*quantity*): 50; number of peers contacted in a swarm for bitfield acquisition (*threshold*): 10; interval between snapshots of a swarm (*frequency*): 8 hours; waiting time between iterations of the algorithm (*interval*): 2 minutes.

# 4 PRODUCERS OF ILLEGAL COPIES

In order to avoid spurious data in our results, three filtering processes were employed to the 58,633 monitored torrents. First we removed all torrents for which all referenced trackers were inaccessible due to malformed URLs. This procedure filtered out 19,490 torrents. Next, we removed all torrents whose swarms could be contacted only in the first iteration of their monitoring. We observed that these swarms could not be further contacted because they were removed from the community. We assume that torrents which are almost immediately (*i.e.* under 8h) removed from the community by the administrators are invalid ones or contain polluted content. This filtering step eliminated 20,027 torrents. Finally, we removed all torrents whose trackers returned error messages upon contact. This final step filtered out 3,001 torrents. To the best of our knowledge, only one of the previous studies related to ours (CUEVAS et al., 2010) employed a similar filtering process and it was not thoroughly explained. Studies that do not properly filter the raw captured data possibly generate biased results with influences from spurious traces.

After the filtering process, there remained 16,115 torrents for investigation. This chapter exposes four main analyses. First we present the characteristics of *producers*, who generate the illegal copies that will be distributed, and *publishers*, who make available the illegal copies in the PirateBay community. We focus on identifying their activity degree and possible relationships among them. Next, we present the most common digitalization processes applied to the observed files, demonstrating their influence in the publishing of copies through the lifetime of a movie. Third, we characterize the first seeders, who bootstrap the dissemination process, acting as initial content providers. Finally, we look into possible relationships among the activities of producers, publishers and providers. The characterization of dissemination from the consumers perspective will be presented in the next chapter.

## 4.1 Digitalization Responsibles

As mentioned in Section 2.1, digitalization groups are responsible for the creation of illegal copies of the movies shared in BT networks. In this study, as previously noted, they are named producers. From the 16,115 analyzed torrents, 10,615 (65.87%) identified the producer that created the file. These copies were created by 623 distinct producers; the 10 most frequent ones are presented in Table 4.1[1]. Figure 4.1 presents a Cumulative Density Function (CDF) in which the horizontal axis represents the producers ordered by volume of created copies and the vertical axis, the cumulative proportion of created torrents. This CDF shows that a small number of producers are responsible for most of the created files: almost 77% of the copies were created by 100 producers (16.05% of the 623 producers).

Table 4.1: Content producers ranking

| **Group** | **Torrents** | |
|---|---|---|
| | # | % |
| Dmt | 388 | 3.66 |
| Cm8 | 300 | 2.83 |
| Mr_Keff | 251 | 2.36 |
| Mastitorrents | 238 | 2.24 |
| Vip3r | 231 | 2.18 |
| Imagine | 215 | 2.03 |
| Ddr | 213 | 2.01 |
| Dutchreleaseteam | 201 | 1.89 |
| Mtr | 178 | 1.68 |
| Extratorrent | 173 | 1.63 |

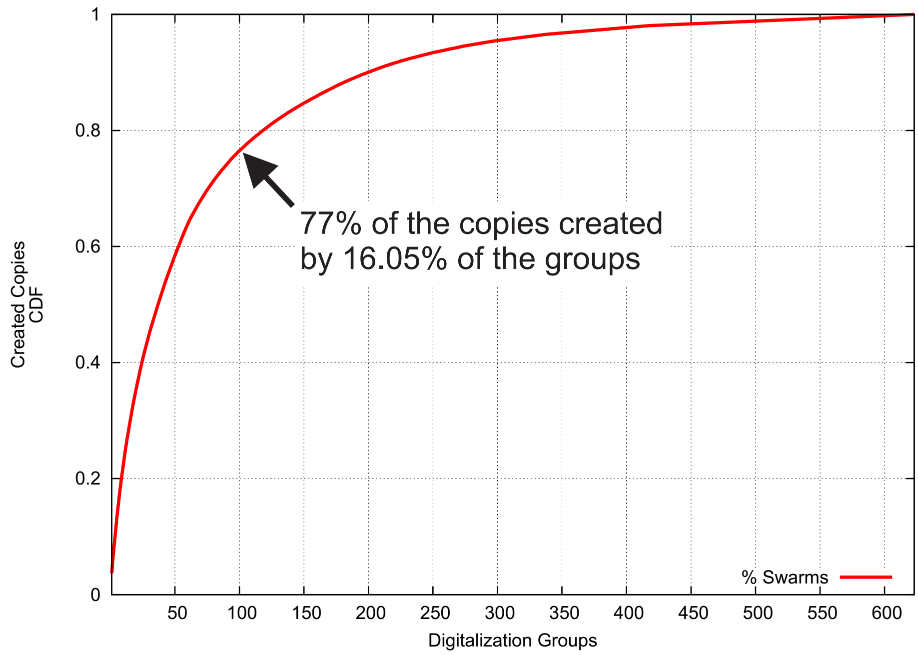



Figure 4.1: Groups cumulative contribution

After a torrent has been created by a producer, it may be published in a community. This step is executed by publishers. These are, in the scope of this study and as mentioned earlier, registered users from the PirateBay community that uploaded

[1]In our analysis, we present the pseudonyms of digitalization groups and community users to help illustrate our results and insights. It should be noted that we do not try to link these names to users real identities. Furthermore, the presented activity rankings are expected to change over time due to file sharing communities dynamicity. Such rankings, however, do not influence the presented results about the dissemination of illegal copies in file sharing communities.

the analyzed torrent files. This community divides its users in four categories (listed in descending order of privilege): VIP, trusted, helper and regular (initial category of every user). The category of a user allows his/her reputation in the community to be inferred. Regarding the 16,115 analyzed torrents, 15,819 were published by 1,153 distinct users (296 torrents were published by users that did not identified themselves). Assuming that each user holds a single identity, table 4.2 presents the most active ones, their category in the community, and the number and proportion of published torrents. It should be noted that, aside from two regular users, all of the most active publishers listed in Table 4.2 are from categories with elevated privileges in the community. Figure 4.2 illustrates a CDF representing torrent publishers activity. The horizontal axis represents the cumulative number of publishers and the vertical axis the proportion of published torrents. This graph shows that few users are responsible for most of the published content: 125 users (10.8% of 1,153) published 81.07% of the content.

Table 4.2: Publishers ranking

| **User** | **Category** | **Torrents** | |
|---|---|---|---|
| | | # | % |
| .BONE. | VIP | 1382 | 8.74 |
| sceneline | VIP | 1377 | 8.70 |
| TvTeam | VIP | 1033 | 6.53 |
| UltraTorrents | VIP | 531 | 3.36 |
| HDvideos | Regular | 368 | 2.33 |
| Black1000 | Regular | 304 | 1.92 |
| SaM | VIP | 299 | 1.89 |
| MeMar | VIP | 284 | 1.80 |
| Sir_TankaLot | Trusted | 269 | 1.70 |
| furtaperas | VIP | 259 | 1.64 |

Figure 4.3 presents the relationship among producers (groups) and publishers (logins). The size of each circle illustrates the number of copies from a specific group that were published by one specific user of the community. Four typical cases were observed: (*i*) strong correlations between a producer and a publisher, as exemplified by case A, which illustrates the user "MeMar" publishing 257 of the 388 torrents produced by the group "Dmt"; (*ii*) a very active user publishing torrents from many groups, as exemplified by case B, in which user ".Bone." published torrents of 227 different groups; (*iii*) a producer supported by different publishers, as exemplified by case C, in which group "Axxo" has its copies published by users ".Bone." and "Test_Verify"; and (*iv*) digitalization group pseudonyms employed as community logins, whose are responsible for publishing a large number of torrents from the group of same name, as observed in D.

## 4.2 Employed Digitalization Processes

Recall from Chapter 2 that distinct digital copies of a movie may have different qualities depending on the employed digitalization process. From the 16,115 analyzed torrents, 13,141 (81.54%) identified the process employed in the creation of the

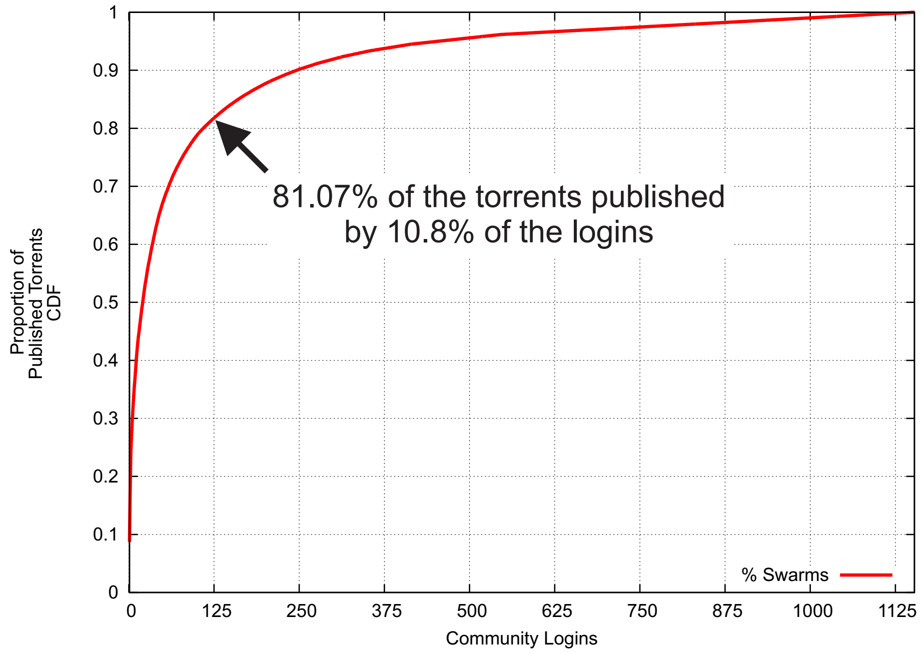



Figure 4.2: Publishers cumulative contribution

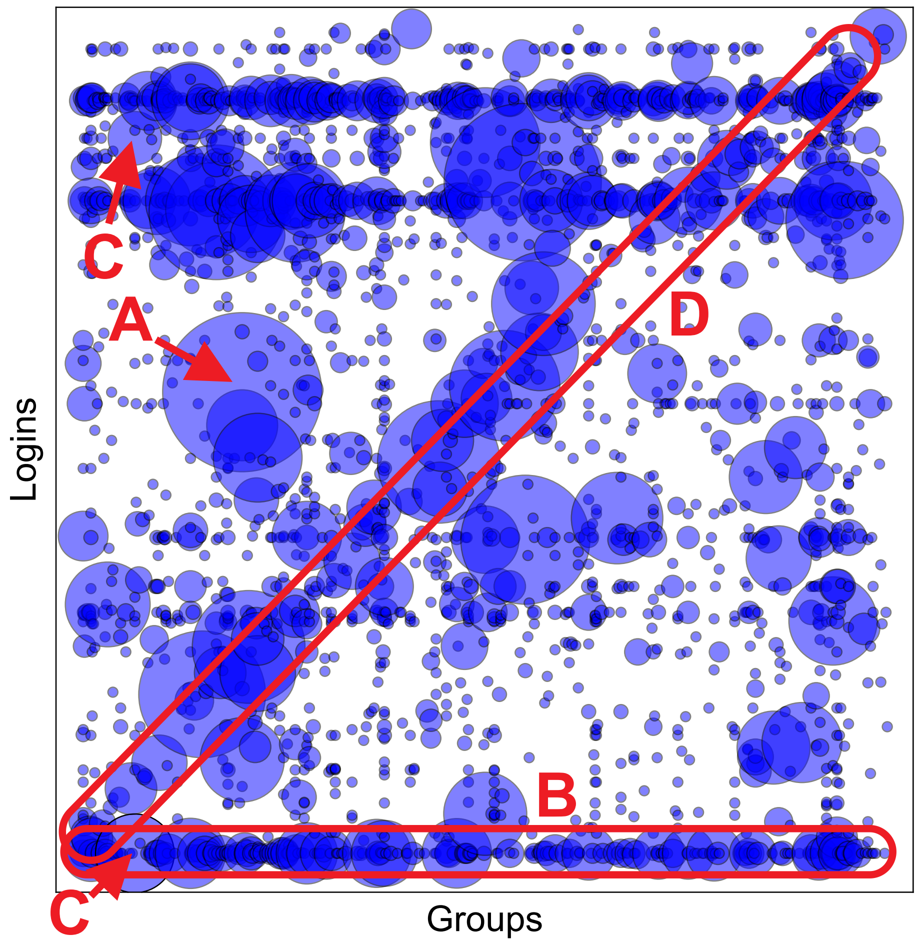



Figure 4.3: Relationships between producers (groups) and publishers (logins)

copy. Table 4.3 presents a correlation between processes and groups, summarizing the processes, their degree of occurrence and the digitalization groups responsible for the greater number of copies within each process.

Table 4.3: Most frequent digitalization processes

| **Process** | **Torrents** # | **Torrents** % | **Main Groups** |
|---|---|---|---|
| DVDRip | 10,623 | 80.83 | Dmt, Mr_Keff, Vip3r |
| TS | 711 | 5.41 | Imagine, Dtrg, Feel-Free |
| R5 | 536 | 4.07 | Cm8, Imagine, Visualise |
| DVDScr | 498 | 3.78 | Mastitorrents, Ddr, Mtr |
| CAM | 372 | 2.83 | Imagine, Feel-Free, Wbz |
| TC | 142 | 1.08 | Mtr, Samurai, Mastitorrents |
| PPVRip | 137 | 1.04 | Iflix, Love, Imagine |
| SCR | 122 | 0.92 | Scr0n, Unkown, Bida |

Results show that the "DVDRip" process is by far the most common digitalization method, being employed in 80.83% of the copies. This prevalence may be explained by two factors. First, the quality of the copies generated using "DVDRip" process present the best quality among the considered types, so it is intuitive that it can attract more interest from users. Second, the DVDRip digitalization method is comparatively simpler and the media (DVD or Blu-ray discs) necessary for the process is widely accessible to the average user.

Other processes that stand out are "CAM", "TS", "DVDScr" and "R5". Their popularity reflects a trade-off among the difficulty of access to the source media for digitalization, the quality of this source and the time after the movie premiere in which it will be available. For example, we observed that the "TC" process presents the best quality among methods with releases expected within the first 4 weeks after the movie premiere. However, since processes "CAM" and "TS" require easily-accessible sources, they are more frequently employed (8.24% in comparison to 1.08% of the "TC" process). Another observed behavior is the specialization of certain groups in the creation of copies that employ processes based on sources that are hard to obtain. For example, the "Imagine" group is the sixth in number of created copies according to Table 4.1. It is, however, one of the main producer of copies based on "TS", "R5" and "CAM" processes. This demonstrates the resourcefulness (and importance) of "Imagine" due to its early access to sources not easily obtainable.

The digitalization processes employed in the creation of copies can also be correlated to the lifecycle of movies after they premiered. Figure 4.4 illustrates the occurrence of copies of nine different movies, produced using various digitalization processes. Circle sizes represent the amount of torrents for each identified copy. The horizontal axis represents, in weeks, how long it took for the movie to be published after its premiere (according to IMDB (IMDB, 2011)).

Three aspects should be noted in the analysis of Figure 4.4. First, the publication of copies using a certain process tends to be concentrated in time, like a burst. This can be observed in the fifth week, when several copies of the movie "Rio" created using the "R5" process arise. Second, torrents containing a copy of

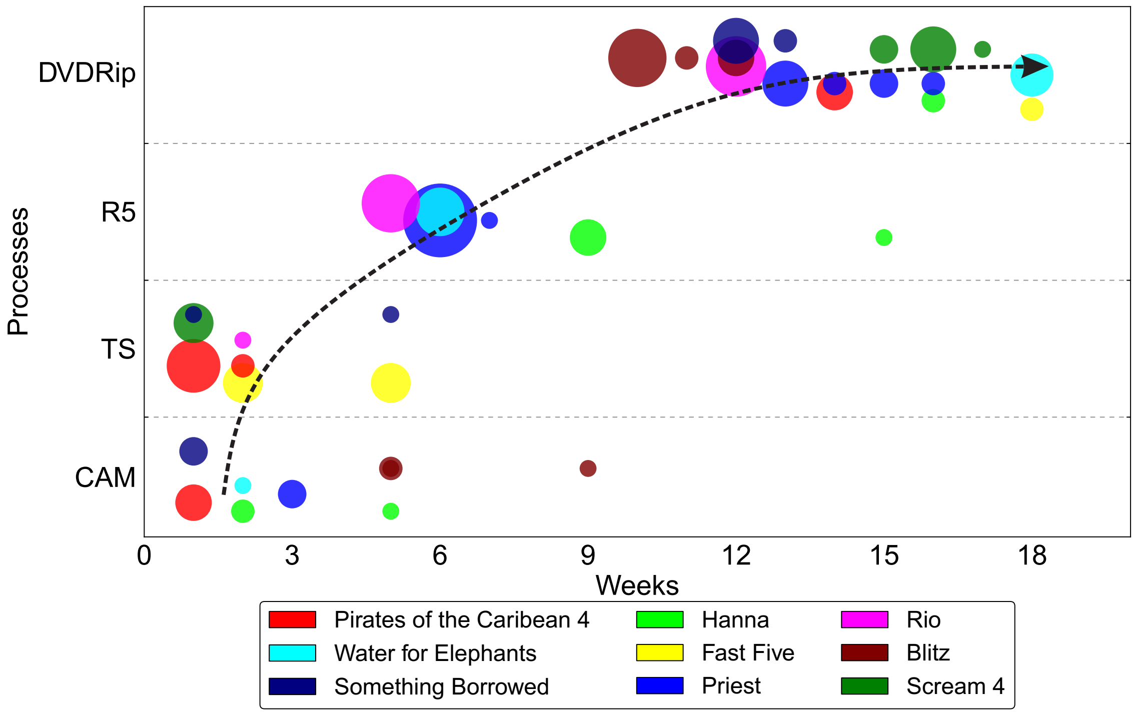



Figure 4.4: Digitalization processes according to weeks after premiere

a movie created using a specific digitalization process may be published even after the initial burst, as observed for the movie "Priest" over the 14th, 15th and 16th week. This behavior seems to occur due to specializations over previously published copies (such as employment of other codec types or the addition of new audio or subtitle languages). The third aspect is that some movies may not appear in all digital formats because of the absence of the corresponding source. Examples of this behavior are observed in "Fast Five" and "Pirates of the Caribbean 4", which did not have copies created with the "R5" process due to, for example, the lack of a DVD region 5 source for these movies.

The data presented in Figure 4.4 was employed to generate the first approximation of a Markov chain representing the evolution of digitalization processes throughout a lifetime of a movie. The resulting model is illustrated in Figure 4.5. Three aspects can be highlighted: (*i*) in the beginning, "CAM" and "TS" processes are typically employed in the creation of the first copies, with the former being predominant; (*ii*) the recurrence encountered in the "CAM" state indicates that the process may be repeatedly employed, for example when a source of better quality is obtained by producers; and (*iii*) the digitalization process of a given movie will eventually converge to "DVDRip", in which case other processes are not used in new copies.

## 4.3 Providers of Illegal Copies

The dissemination of illegal copies in BT networks depends on providers. These are, as previously mentioned, users with the first copy of the file. Without these

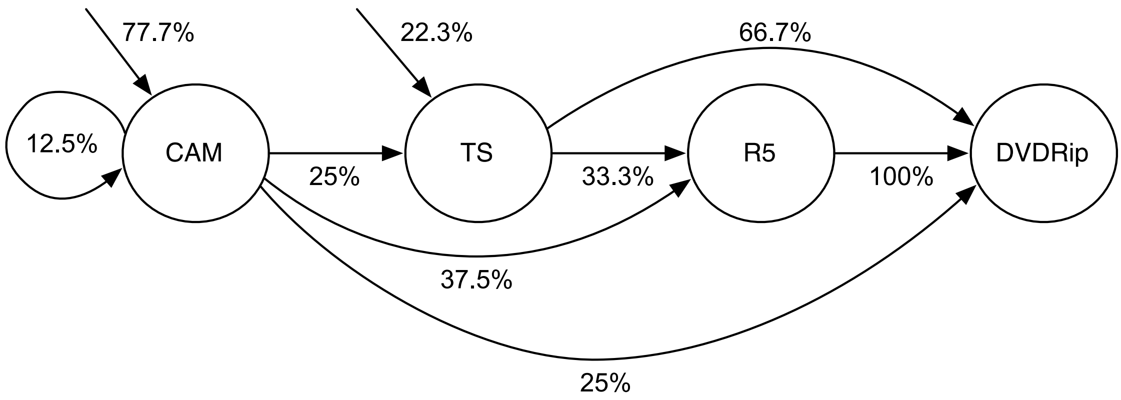



Figure 4.5: Markov chain of digitalization processes evolution

providers, dissemination would not be possible. Recall that, in BitTorrent, the user who publishes the torrent in the community is not necessarily the one in possession of the file containing the copied content. These users, also known as first seeders, are responsible for bootstrapping the swarm. To characterize these users, we analyzed the peerlist obtained from trackers early in swarms lifetime.

Torrents added to the community were captured as soon as they were detected by the TorrentU community lens (recall from Chapter 3 that the community is probed in intervals of 2 minutes). Once detected, the corresponding torrent trackers were contacted for receipt of the initial peerlist. The shorter the time between a torrent publication and receipt of its peerlist, the higher the chances that only the first seeders will be contained in the tracker response. In our study this interval was 4 minutes on average.

We begin our analysis by characterizing each contacted tracker. From 1,098 observed trackers, 181 (16.48%) answered the query with a valid peerlist. We were able to identify the geographical location of 164 trackers. Figure 4.6 illustrates the obtained results. In the graph, each bar represents the number of trackers found in a specific country and its color the respective continent. Results indicate that Europe stands out as the location of most trackers with a total of 76 hosts, 40 of which are on the Netherlands. North America appears in the second position, mainly due to the United States, which hosts 37 trackers. Finally, Asia appears as the third continent in number of trackers. These specific results are overall in line with other general ones previously obtained by Zhang *et al.* (ZHANG et al., 2010).

After contacting the trackers, the obtained responses were analyzed and led to the identification of the first seeders in 7,692 (47.73%) of the torrents. These were seeded by 9,254 peers (a few swarms presented more than one initial seeder). These peers were associated to 2,810 unique IP addresses. Table 4.4 presents the list of most active users, indicating their country of origin, ISP, and number of swarms joined as initial seeder. Figure 4.7, in turn, show a CDF of the cumulative contribution of first seeders according to the proportion of analyzed torrents.

From Table 4.4 and Figure 4.7, two insights can be obtained from the analyzed results. The first is that 8.89% of the observed IP addresses (250 of 2,810) par-

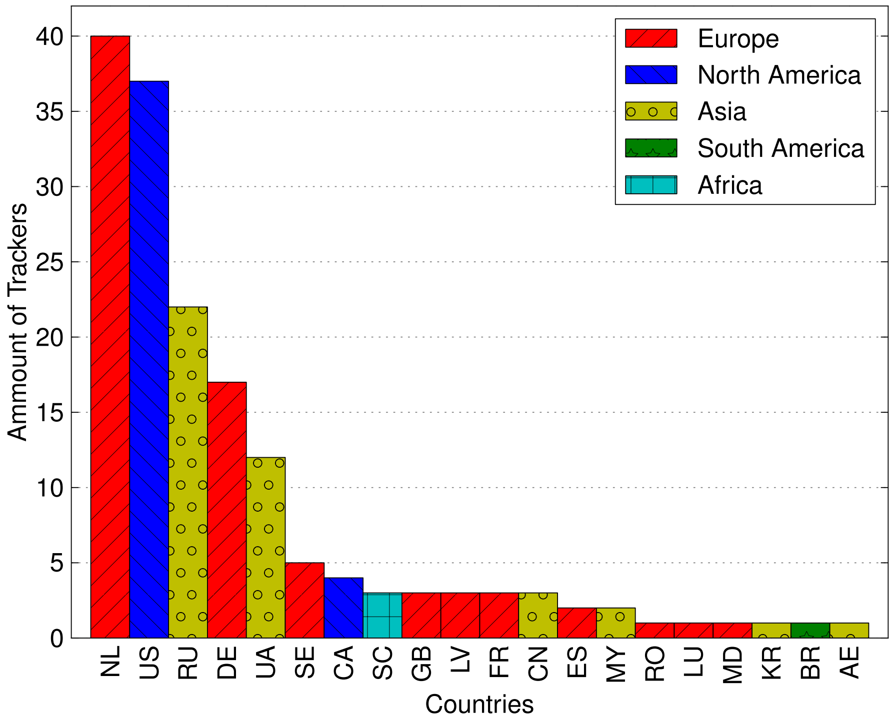



Figure 4.6: Geographical location of trackers

ticipated as first seeders of 67.52% of the analyzed swarms. This result indicates the possibility that specialized users are employing *seedboxes* (WIKIPEDIA, 2011b) in order to disseminate their content. The second aspect is that 87.58% of the IP addresses (2,461 of the 2,810) exclusively seeded one or two swarms. Such behavior indicates that these users may be "domestic" ones, sharing illegal copies of very specific content types.

Table 4.4: Seeder ranking

| Country | ISP | # Swarms |
|---|---|---|
| US | - | 506 |
| US | - | 503 |
| FR | Ovh | 476 |
| FR | Ovh | 243 |
| NZ | Obtrix | 186 |
| ES | - | 124 |
| FR | Ovh | 107 |
| NL | Ziggo | 104 |
| FR | Ovh | 97 |
| PL | Mokadi | 94 |

In order to identify the location of the observed content providers, we considered the first seeder of each swarm a unique entity, even if peers from distinct swarms presented the same IP address. We successfully determined the location of 9,051 of the 9,254 observed seeders. Figure 4.8 shows the 30 countries presenting higher concentration of first seeders (countries not presented in the graph always contained less than 20 seeders). Results indicate that Europe stands out as the most common

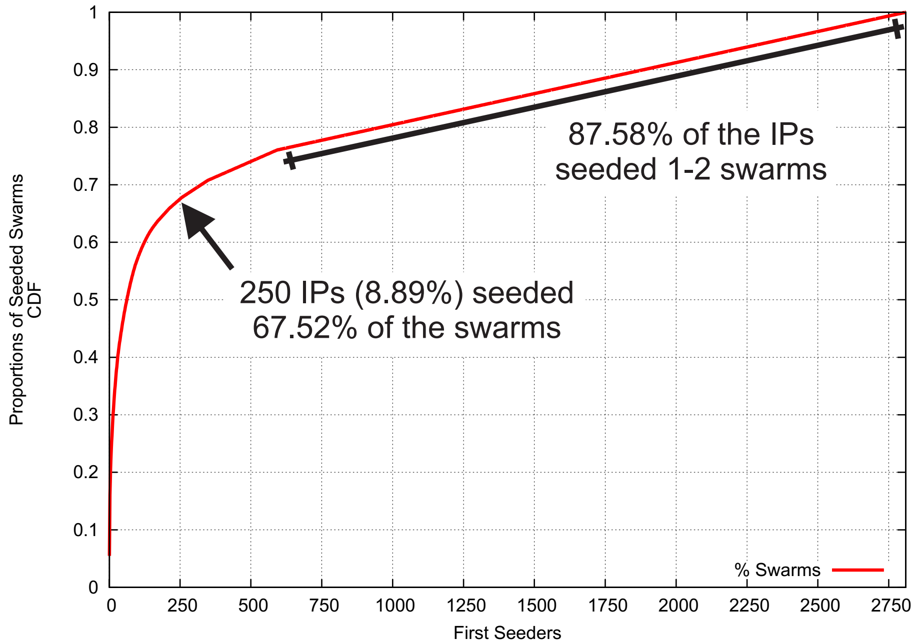



Figure 4.7: Providers cumulative contribution

location of content providers, hosting nearly four times more first seeders (59.45%) than North America (18.39%), which appears in second place. Another behavior observed is that France, United States and the Netherlands present considerable higher number of first seeders than the average measured for the other countries.

Through the characterization of providers (first seeders), producers (digitalization groups) and publishers (community users) we identified examples that indicate the existence of relationships among these entities. Figure 4.9(a) presents the correlation of digitalization groups and first seeders. Three points of interest are highlighted: First, some providers are dedicated to the dissemination of specific producers copies. This can be observed in A, which represents the group “Miguel”, which had 91.04% of its swarms seeded by two IP addresses only. Second, some providers serve copies of various producers, as observed in case B. Third, a producer may be served by a diverse group of providers. This may be observed in case C, which represents the group “Dmt”, which had its copies provided by 83 different seeders.

Figure 4.9(b) characterizes associations between community logins and first seeders. Three points of interest are highlighted. First, providers with high degree of activity may be associated with a single community login. This can be exemplified by case A, in which one specific IP seeded copies published by user “TvTeam”. Second, one provider may serve a diverse group of publishers, as in case B. Finally, one publisher may be served by a diverse group of providers. This corresponds to case C in the figure, representing a user (“MeMar”) who had its published copies seeded by 61 unique IPs.

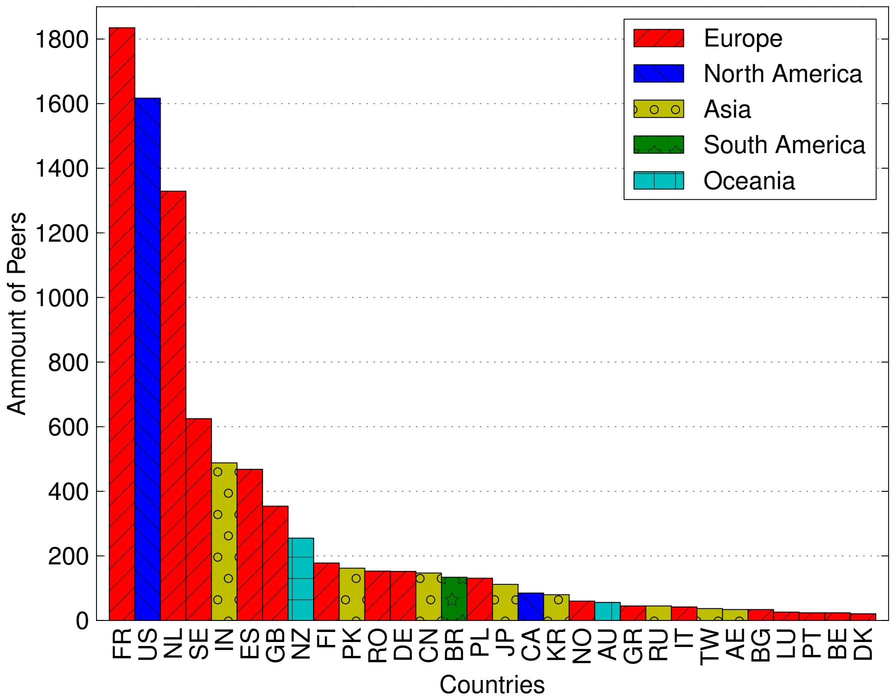



Figure 4.8: First seeders location

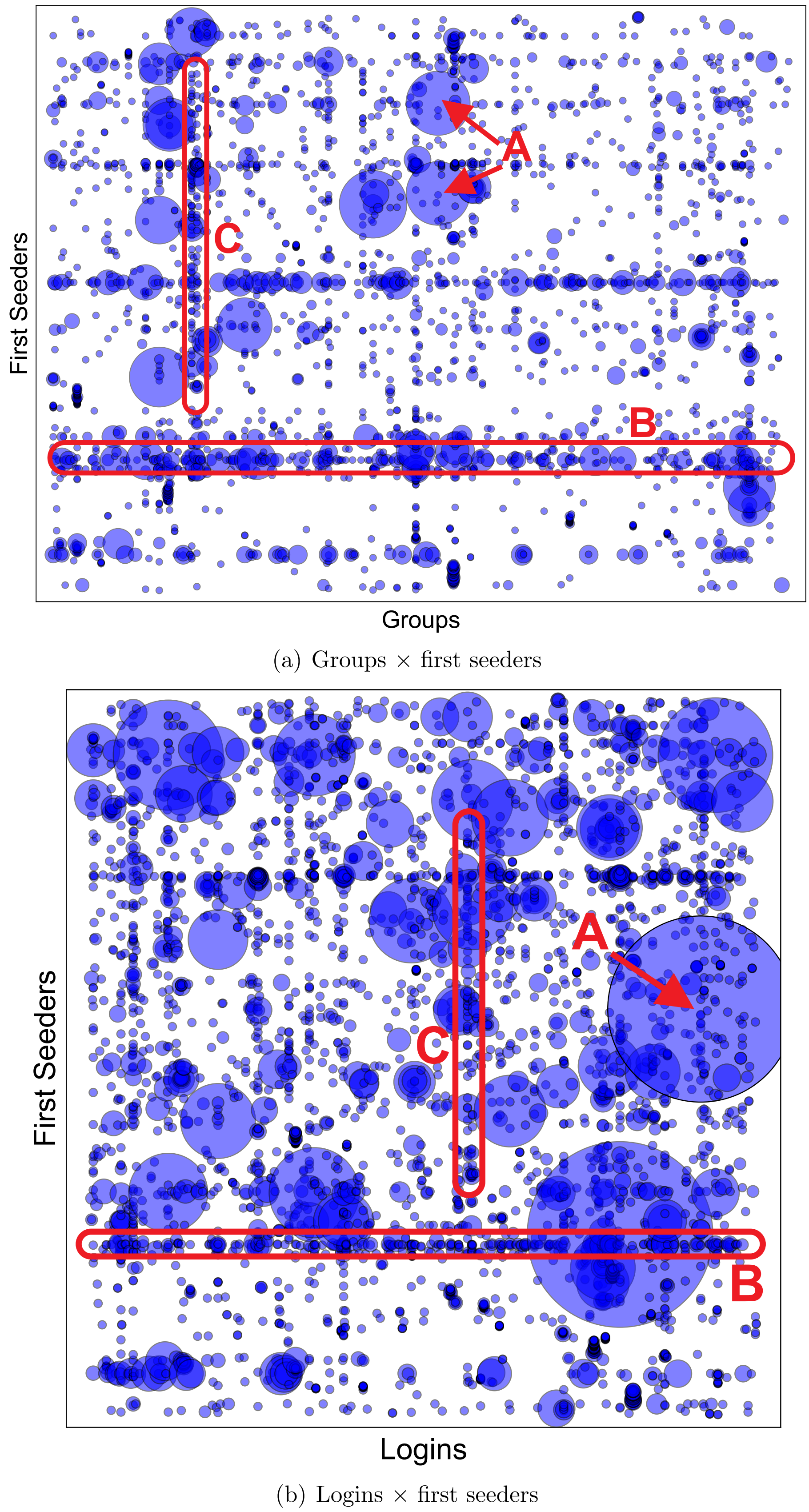



Figure 4.9: Relationships among producers (groups), publishers (logins) and providers (first seeders)

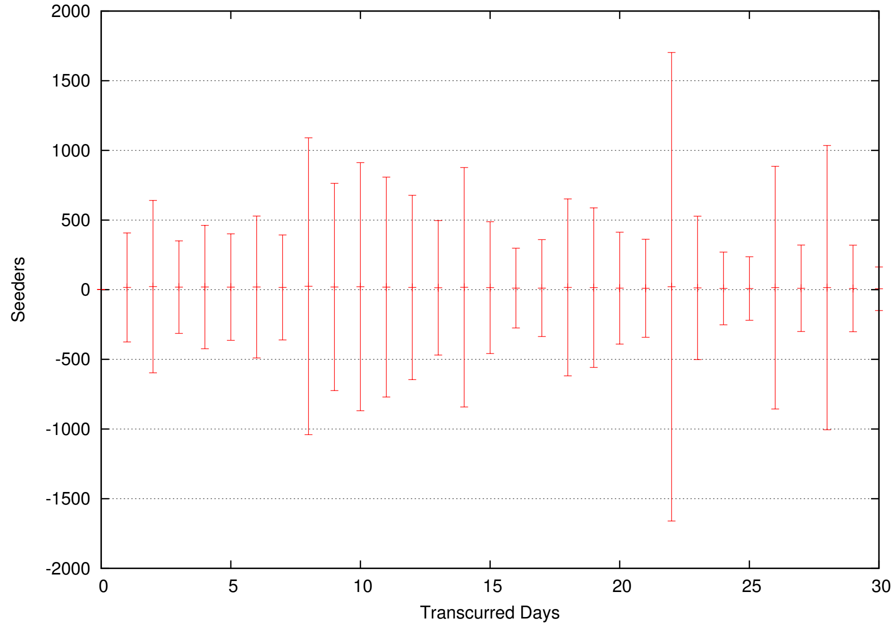
# 5 CONSUMERS OF ILLEGAL COPIES

As mentioned in Chapter 3, our monitoring infrastructure kept track of the development of all swarms that had its first seeder(s) identified. During this process, we were able to catalogue development characteristic of 789 swarms, totalling an amount of 923,009 distinct IP addresses participating at a certain moment throughout our monitoring. The presentation of the obtained results is divided in two sections: the first emphasizes on aspects regarding the development of swarms while the other presents observations concerning the characterization of peers that consumed these contents. To be more precise, we initially present the peculiarities involving seeder and leecher dynamics throughout our monitoring lifetime. Next, we characterize swarm development taking into account influencing factors, such as digitalization groups and processes. Entering the following phase, we first observe and determine the level of participation of each peer in different swarms and then characterize these consuming peers giving a proper emphasis on those that presented a high level of participation throughout our monitoring period. This last set of frequently participant peers will be referred throughout this chapter as “big downloaders”.

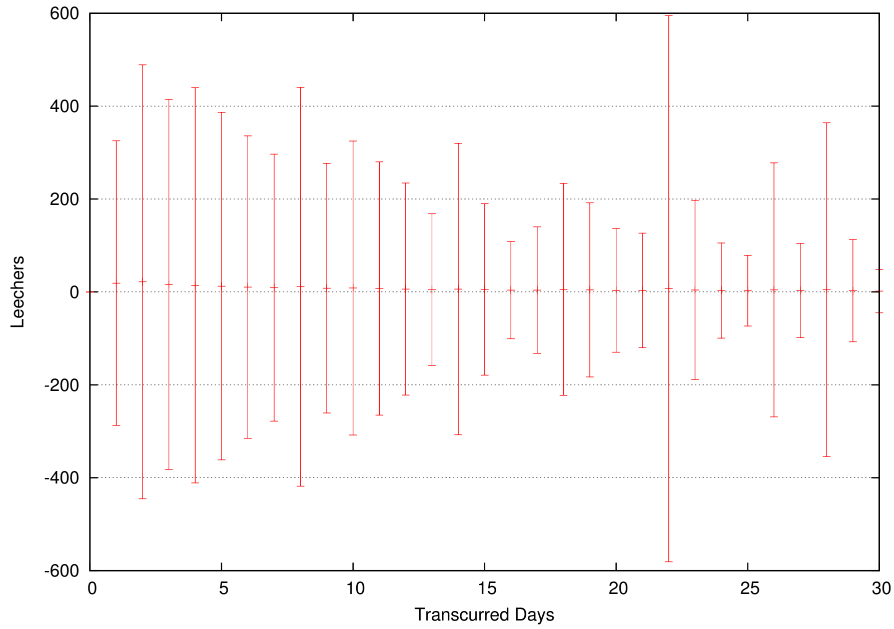
## 5.1 Consuming Dynamics

In order to characterize the development of a swarm, we account the variation in the number of seeders and leechers over an observed timeline. For us to present a clear and easily understandable set of results, we processed the information by firstly stipulating a day as the timeline tick for observational purposes. From that on, the amount of seeders and leechers catalogued was compiled into a single daily value, by calculation of simple average of participants per day. Since an average can sometimes disguise several characteristics of an obtained result, we have also determined each day’s standard deviation for all swarms.

A general view of the obtained results is presented in Figure 5.1 (a and b). These figures characterize average swarm development according to the amount of seeders and leechers, respectively. Standard deviation values include some unusual results, such as a negative amount of participants. This occurs due to a high level of variation in our traces. However, it would be a misconception to consider that a trustworthy view can be achieved by removing these outliers. BT swarm development differ emphatically over one of its main aspects: the popularity of the content being shared. Therefore, for us to properly visualize the information hidden in our traces, the result set was clustered.




(a) Seeder development




(b) Leecher development

Figure 5.1: Mean and standard deviations characterizing swarm development

In order to establish tiers separating swarms according to content popularity, we observed the obtained results and divided the swarms based on their peeks of participants (seeders plus leechers) throughout the 30 days of monitoring. The clusters were determined as follows: swarms that reached a maximum of 50 peers; swarms that ranged from 51 to 100 peers; swarms that ranged from 101 to 500 peers;

and swarms that surpassed 500 peers. The results obtained for the first cluster are not presented because they are similar to those illustrated in Figure 5.1 (a and b). The majority of swarms in this cluster remained unpopulated by peers during the monitoring process, which caused the mean values to be pulled downward and the distribution to be non-normal. On the other hand, the following clusters' swarms presented a more homogeneous behaviour.

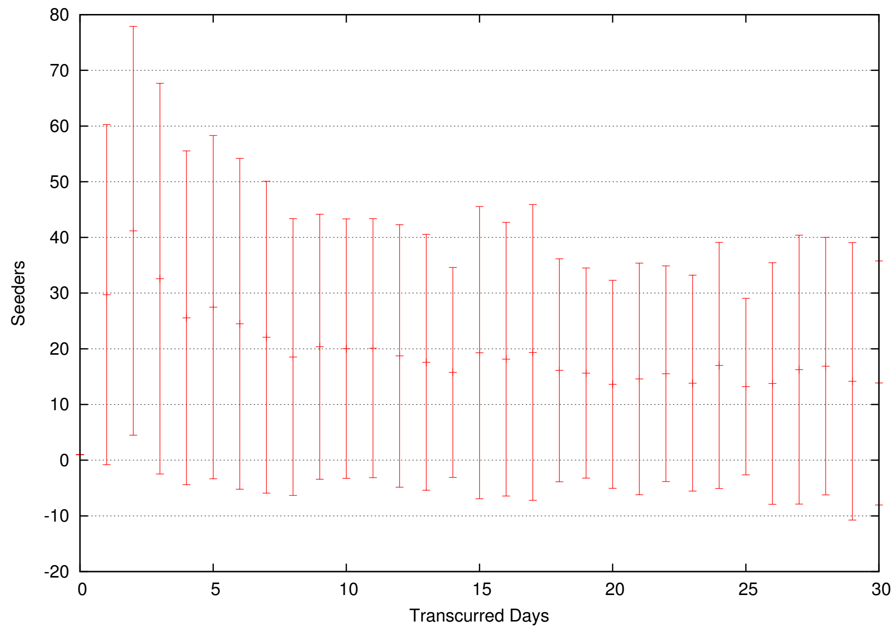
Observing the second cluster in in Figure 5.2 (a and b), one can perceive a slight increase of seeders during the first three days of existence followed by a stabilization from the 10th to 30th day at the range of average 14 - 17 seeders. Likewise, the amount of leechers experience an initial burst (with greater intensity) during the first three days. This is followed by a deflating period from the 4th to the 10th and a stabilization at a range of average 8 - 12 leechers until the end of the month. Interestingly, the second and third clusters presented similar behaviour as one can observe comparing Figures 5.2 and 5.3. The fluctuations describing the variation on the mean average of seeder/leecher as well as the standard deviation are proportionally equivalent. This allows us to believe that there is a group of averagely popular contents that lead to similar swarm development patterns, such as the one described at the beginning of this paragraph.

The results obtained for the fourth cluster of swarms are presented in Figure 5.4 (a and b). The fluctuation of the mean values over the days diverges substantially from the one observed for the previous clusters, which allow us to speculate that this cluster represents a different behavioural pattern. Unlike what has been observed at the second and third clusters, the mean values over the days can not be clearly characterized into an inflating, deflating and stabilization period. These swarms' development behave in the form of a wave with decreasing strength that periodically increases the amount of participants, as it can be observed at days 8, 14 and 18. Also, due note that the unexpected peaks that occurred at days 22, 26 and 28 are just consequences of sudden surges of interest for few extremely popular swarms.

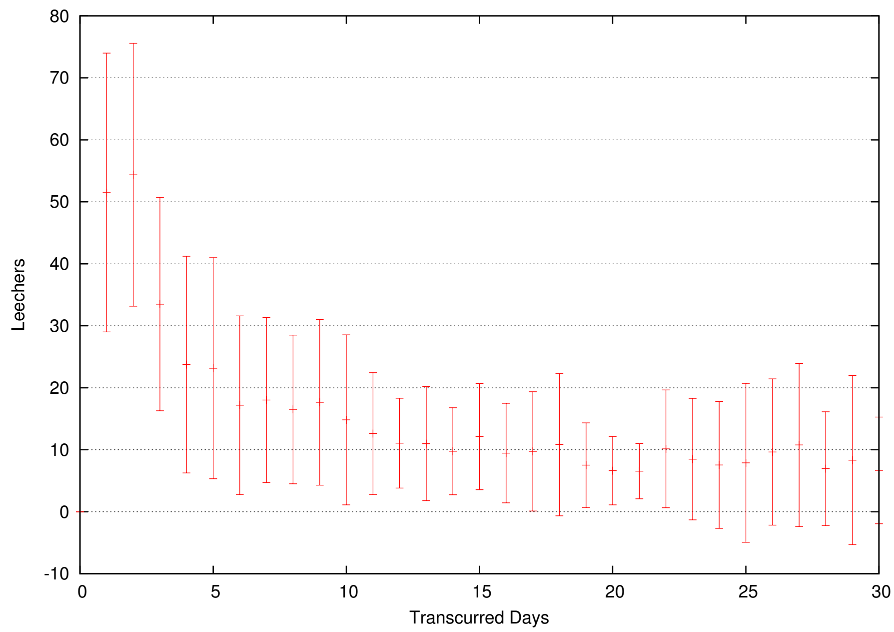
The analysis just presented revealed swarm development patterns considering different swarm sizes. Now we report results obtained by evaluating swarm development as a function of digitalization groups and processes. As far as we are aware of, such an analysis was not performed in previous work and, therefore, is an important contribution of this work.

Initially, we analyze swarm development considering the set of swarms bootstrapped by 441 torrents that properly identified the digitalization groups and the set of swarms whose 348 torrents did not mention any digitalization group. As one can observe in Figure 5.5 (a and b), the impact of this identification on swarm popularity (and, indirectly, "health") is quite evident. Quantifying this heterogeneity, swarms from torrents that identified the responsible digitalization group have averagely 6.48 times more seeders and 4.41 leechers then those that didn't.

Now we analyze swarm development according to identification (or not) of digitalization processes (in corresponding torrents). Figure 5.6 (a and b) illustrates the results. As one can easily note, swarms of torrents that identify a process have an overall higher popularity. They presented an average of 2.29 times more seed-

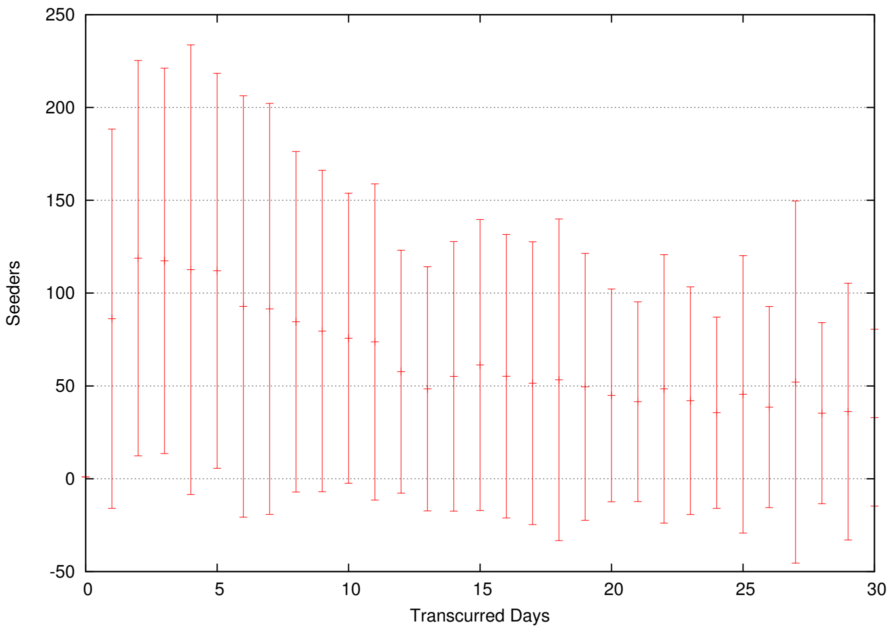



(a) Seeder development

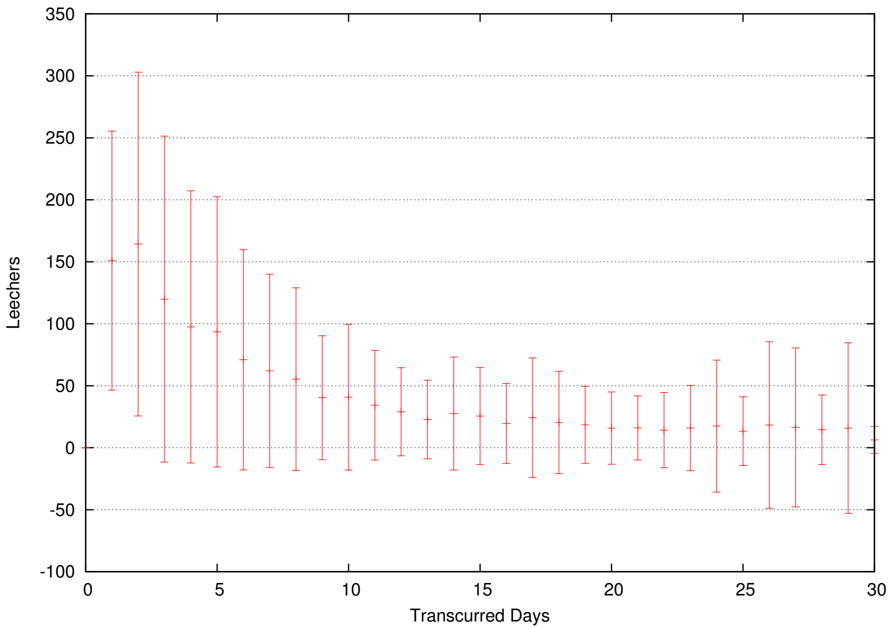



(b) Leecher development

Figure 5.2: Characterization of swarms that ranged from 51 to 100 peers

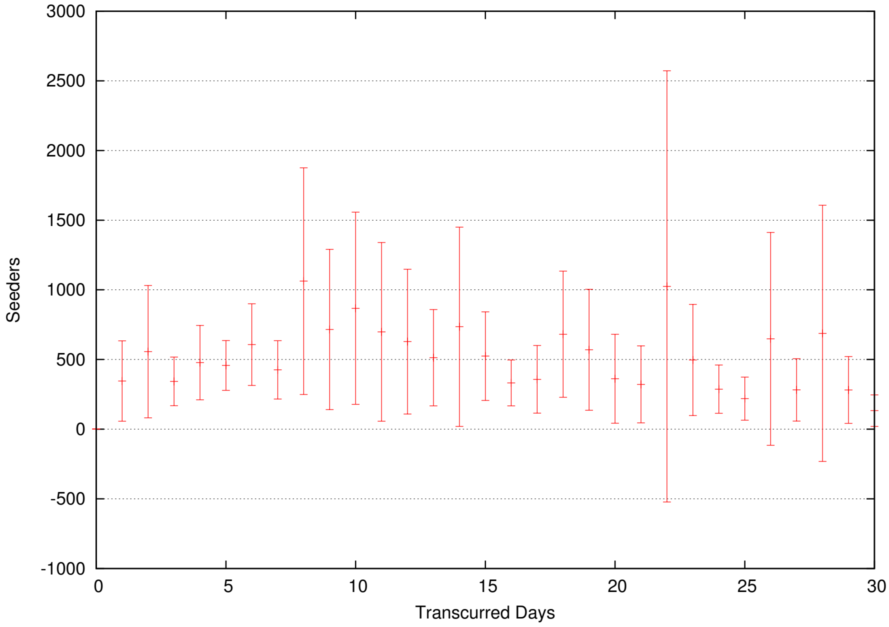



(a) Seeder Development

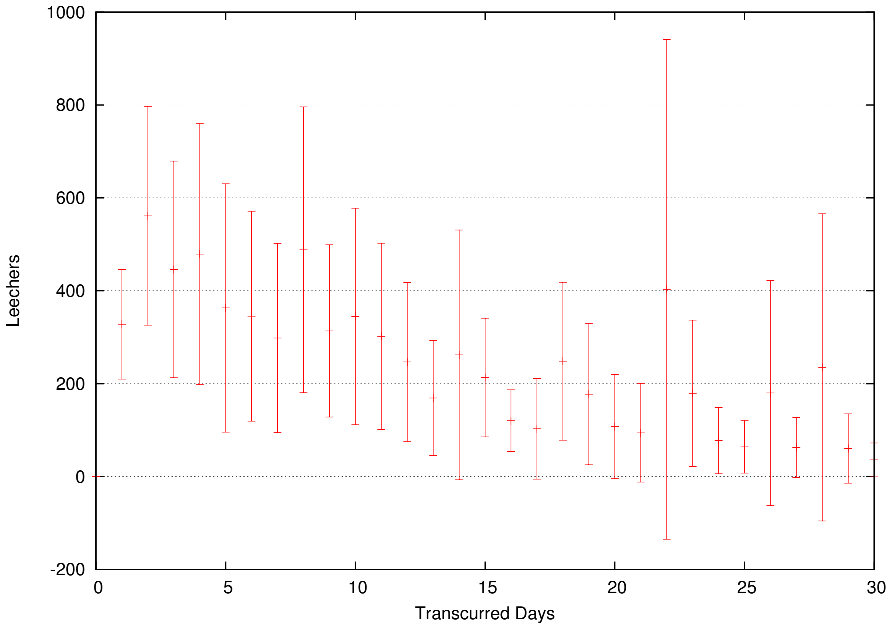



(b) Leecher Development

Figure 5.3: Characterization of swarms that ranged from 101 to 500 peers

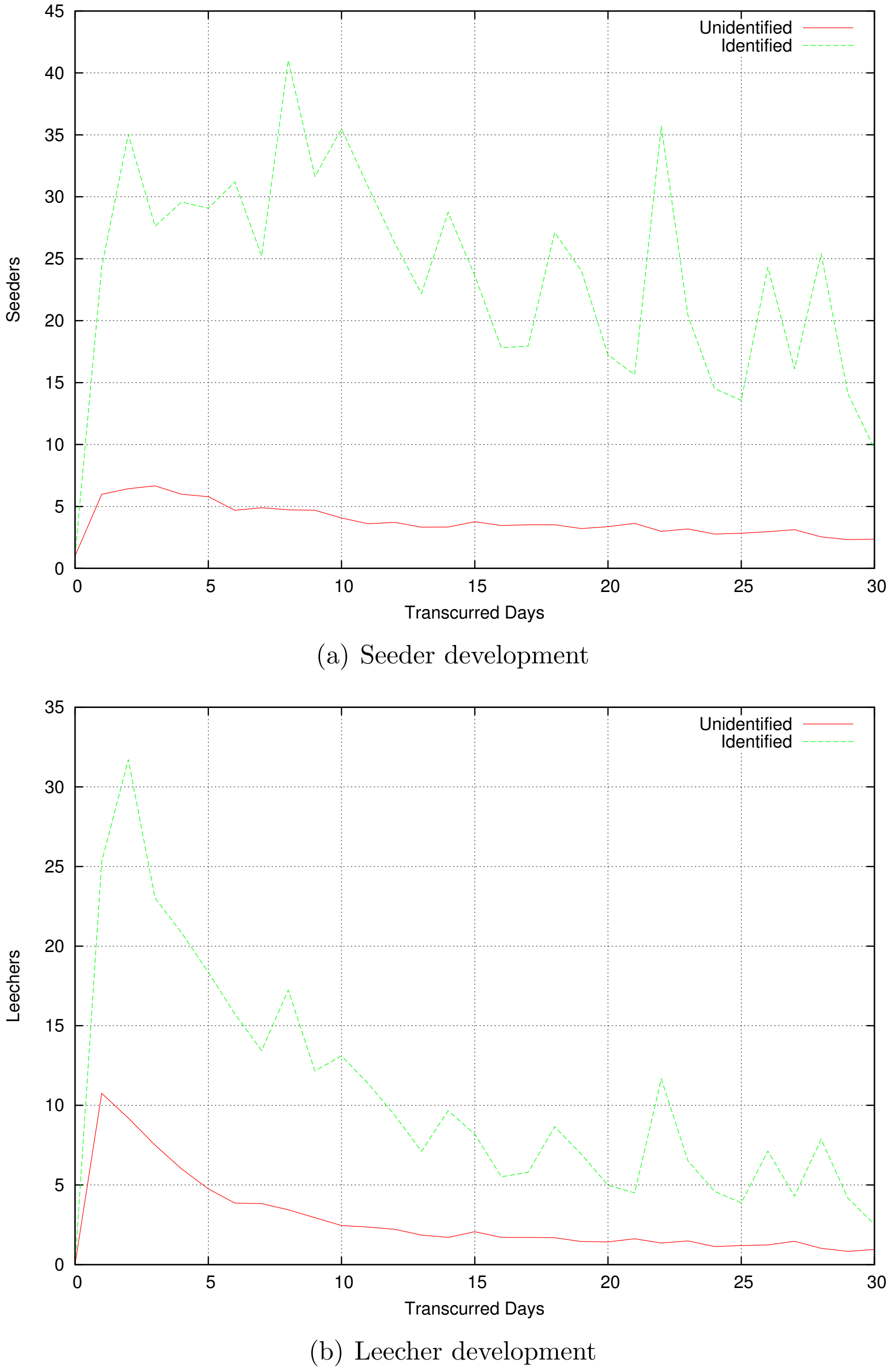



(a) Seeder development




(b) Leecher development

Figure 5.4: Characterization of swarms that surpassed 500 peers




(a) Seeder development



(b) Leecher development

Figure 5.5: Characterization of swarms in respect to digitalization group identification (in corresponding torrents)

ers and 1.85 leechers. Even though these values are lower than the ones presented when analyzing digitalization group identification, they show that torrents whose digitalization process is specified tend to positively influence the popularity and interest in the corresponding swarms. Note that the anomaly observed from days 3 to 7 at Figure 5.6(b), when the amount of leechers is greater at swarms without process identification, happens because the vast majority of contents being shared in this group are of greater appeal. Which causes this group of swarms to have an initial burst leading to a higher peak, followed by a fast descend and, finally, a lower amount of interested peers until the end of the monitoring. The reasons for publishers and providers to disseminate these contents is out of this dissertation scope.

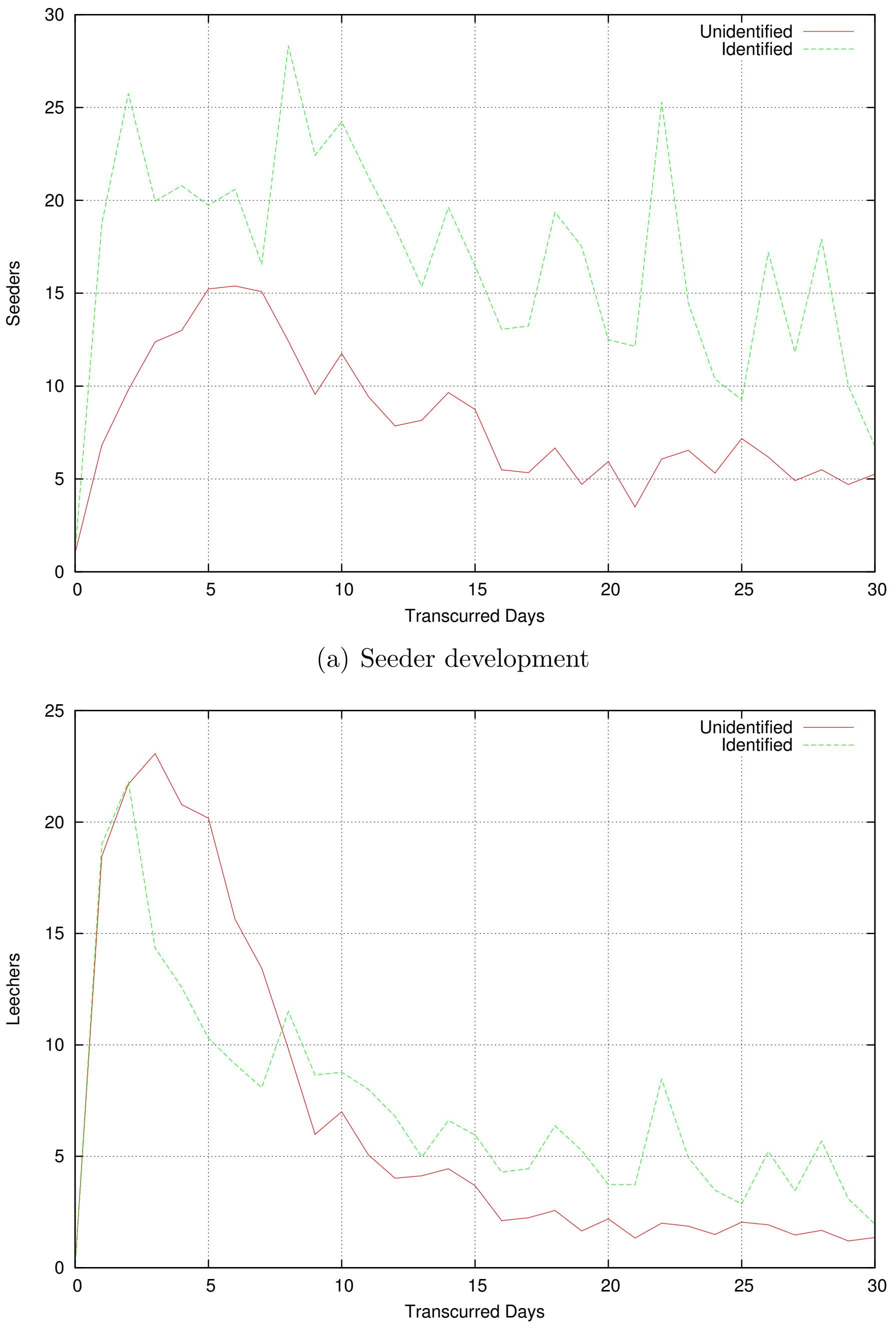
## 5.2 Consumer Characterization

In this section we move from a macroscopical overview of swarm development to a microscopical one, in which we analyse the observed end users over this month's monitoring. Figure 5.7 presents a CDF showing users (IP addresses) and number of swarms of which they participate. It is possible to observe the high activity of a few users and the massive participation of sporadic ones. This observation allows us to speculate that few users are responsible for the majority of content downloads and, from this point on, we will be referring to them as "big downloaders".

Aside from peer participation, another important factor that we analysed is their location. Figure 5.8 shows the 20 countries presenting the highest amount of participating IP addresses. India and Sweden come highest in this ranking. This result might probably be explained by the regional nature of the observed shared contents (such as translated audio tracks, hardcoded foreign language subtitles and movie production origin). To better understand the result, we performed an in depth analysis of every content being shared, searching for at least one of the three cited regional factors. The result was that, out of the 789 movie files, 211 had regional factors pointing to Spain, 150 to Sweden and 71 to India. Due to this high percentage (54.75%) of regionally identified content within our sample, it would be an erroneous assumption to presume that these countries represent the majority of overall infringing copies of content consumers. Since we only monitored swarms that we identified their first seeder(s), it is plausible to infer that the majority of swarms born after their torrent publication at Piratebay are of regionally identified content.

For us to observe characteristics of the "big downloaders", we filtered our by determining that only peers that participated of at least 12 swarms would fall under this category. This filtered view led to a total of 972 peers and Figure 5.9 shows their locations grouped by countries. Spain presented more "big downloaders" then the following four countries put together, followed by Sweden with an amount of peers similar to the one from the Philippines. Once again, the regional factor standing out.

Notice how come India drops from first at Figure 5.8 to fifth at Figure 5.9. That can be explained since the only way for us to determine a peer's identity is looking at




(a) Seeder development



(b) Leecher development

Figure 5.6: Characterization of swarms regarding digitalization process identification

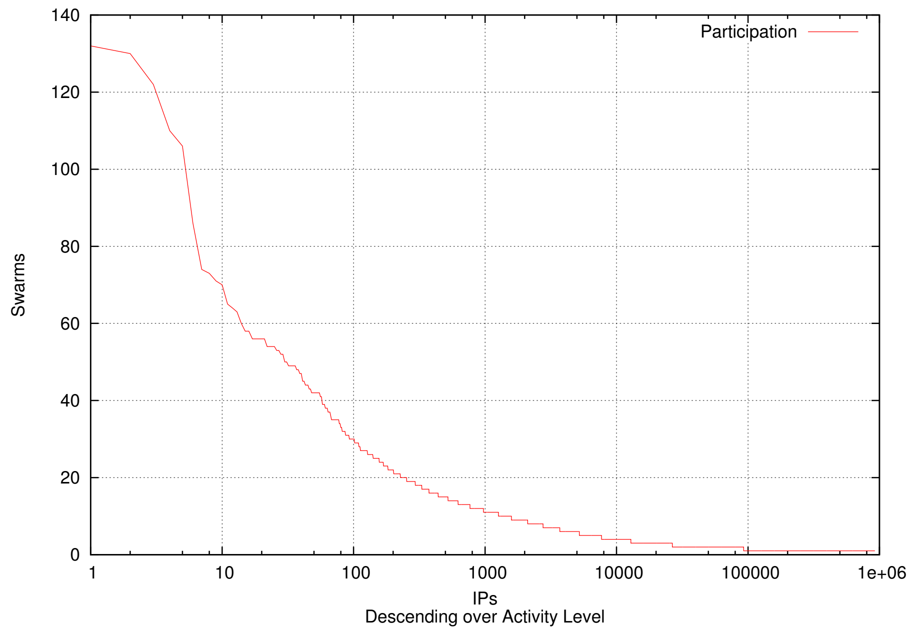



Figure 5.7: Activity level of consumers

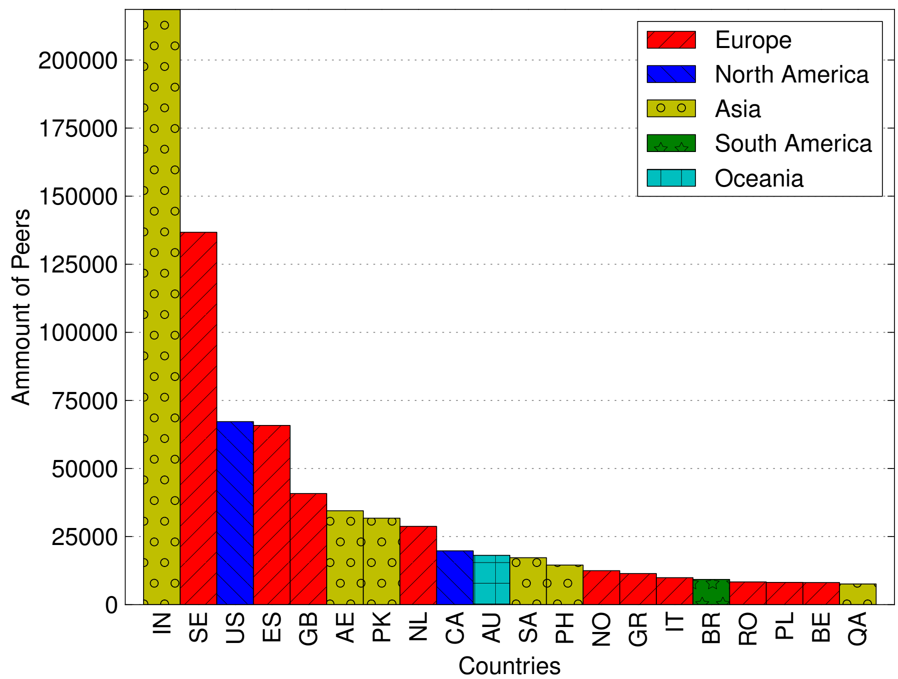



Figure 5.8: Top location of consumers

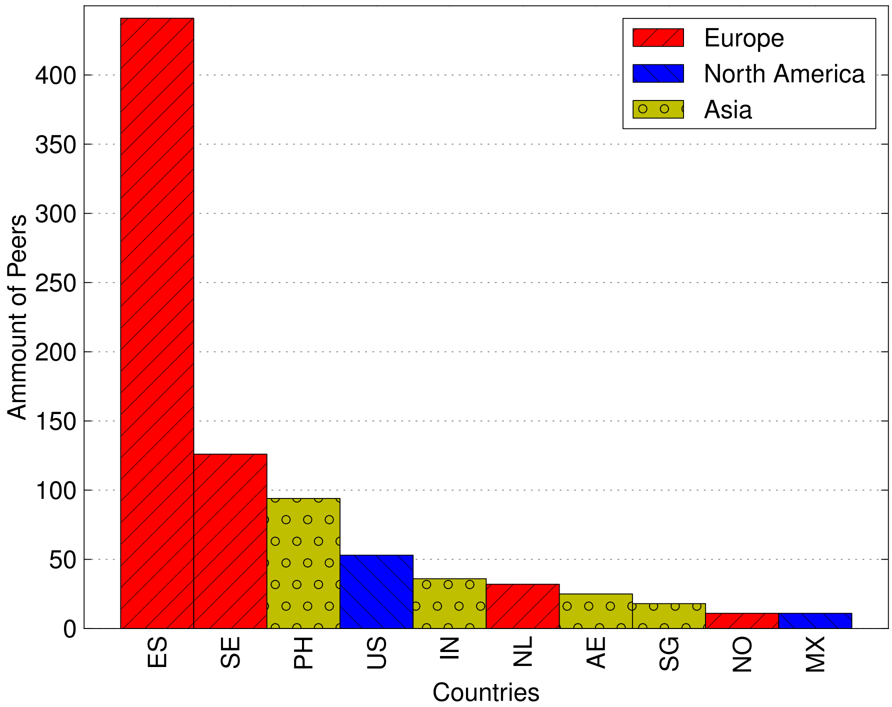



Figure 5.9: Location of big downloaders

its IP address, which could lead to a NAT (or an ISP) with a group of users behind it. Given the nature of IP networks nowadays, we recognize that this possibility might have influenced on our results. Alas, to the best of our knowledge at the present date there is no way to perform a more thorough analysis. Geolocation services might have been more extensively utilized to aid in this identification but even they have flaws of their own and can not be completely trusted.

# 6 CONCLUSIONS

Content sharing through BitTorrent networks is one of the activities that generate most of the Internet traffic. It is known that most of this traffic is related to the sharing of illegal copies of various types of copyrighted content. So far, even with the relevance of this research topic, no studies related to the characterization of BitTorrent content sharing have focused on mapping the dissemination dynamics of illegal copies. Aiming at bridging this gap, we presented a detailed study of traces registering seven months of activities from one of the most, if not the most, popular BitTorrent file sharing community. Traces were collected with an extension of a BitTorrent monitoring architecture designed to observe the BT "universe". The analysis of the collected data allowed us to obtain new insights about the dissemination of illegal copies of content. To the best of our knowledge, this is the first scientific study that focuses on characterizing the dissemination and consumption of illegal copies of content in BitTorrent networks.

Based on the obtained results it is possible to identify behavior patterns of the sources distributing illegal copies of content. Regarding the producers of copies, it was found that most torrents present an identification of a digitalization group responsible for its creation and that most copies are generated by a small number of groups. In the case of publishers, a behavior similar to the one exhibited by producers was observed, in the sense that most torrents are published by a small number of active users. An association between producers and publishers was also identified. Analysing the employed digitalization processes, we discovered that certain producers are specialized in specific processes. Our study helps understanding and quantifying the evolution of the digitalization processes employed throughout the lifetime of a movie after its premiere. By analysing the peers responsible for the initial seeding of illegal copies, we identified relationships among producers, publishers and providers.

After a thorough analysis of the dissemination aspects regarding sharing of illegal copies of content, we focused on the consumption of such media. Initially we have divided the groups of swarms accordingly to their content's popularity so that the impact of this aspect on a swarm's lifetime development could be sought out. Following, the consequences of a digitalization group identification to a swarm development were outline and measured to an expected differential value (comparing swarms that did contain an identification to those that did not). Following the same methodology, the statement of an applied digitalization process was also measured, this time considering the differences among each type of process as well. During these analysis it was able to notice that all of the above factors have a notable influence on a BT swarm lifetime development and, consequentially, it's size expectancy. At the final analysis phase of our methodology, we listed the top location of users consuming these infringing copies of content. We identified regional aspects that affected the observed outcome and, finally, presented an overall view of consumers participation. Concluding this phase we were able to state that not only distributors respect the pattern determining that few are responsible for most, as well as consumers do.

The obtained results are relevant for operators of Internet, media service providers, the film industry as well researchers of this community. With these results we hope that we assist third parties on the development of designing effective models and mechanisms to securely operate large-scale content delivery solutions.

During our study, we identified three opportunities for future work. The first one consists in observing the behavior of users from other vastly utilized BT communities. A second opportunity is the observation of "darknets": private BT communities that might be the starting point for the dissemination of illegal copies of contents later found on public communities. Finally, the opportunity to utilize the same methodology here presented for other types of infringing contents being shared over BT networks.

# REFERENCES

BAUER, K.; MCCOY, D.; GRUNWALD, D.; SICKER, D. BitStalker: accurately and efficiently monitoring bittorrent traffic. In: WIFS 2009, IEEE WORKSHOP ON INFORMATION FORENSICS AND SECURITY. **Proceedings...** [S.l.: s.n.], 2009. p.181–185.

CHOW, K.; CHENG, K.; MAN, L.; LAI, P.; HUI, L.; CHONG, C.; PUN, K.; TSANG, W.; CHAN, H.; YIU, S. BTM - An Automated Rule-based BT Monitoring System for Piracy Detection. In: ICIMP 2007, INTERNATIONAL CONFERENCE ON INTERNET MONITORING AND PROTECTION. **Proceedings...** [S.l.: s.n.], 2007. p.1–6.

CUEVAS, R.; KRYCZKA, M.; CUEVAS, A.; KAUNE, S.; GUERRERO, C.; REJAIE, R. Is content publishing in BitTorrent altruistic or profit-driven? In: CO-NEXT 2010, INTERNATIONAL CONFERENCE ON EMERGING NETWORKING EXPERIMENTS AND TECHNOLOGIES. **Proceedings...** [S.l.: s.n.], 2010. p.1–12.

ENVISIONAL. **An Estimate of Infringing Use of the Internet**. Available at `http://documents.envisional.com/docs/Envisional-Internet_Usage-Jan2011.pdf`, access in July 2011.

IMDB. Available at `http://www.imdb.com`, access in July 2011.

JUNEMANN, K.; ANDELFINGER, P.; DINGER, J.; HARTENSTEIN, H. BitMON: a tool for automated monitoring of the bittorrent dht. In: P2P 2010, IEEE INTERNATIONAL CONFERENCE ON PEER-TO-PEER COMPUTING. **Proceedings...** [S.l.: s.n.], 2010. p.1–2.

LE BLOND, S.; LEGOUT, A.; LEFESSANT, F.; DABBOUS, W.; KAAFAR, M. A. Spying the world from your laptop: identifying and profiling content providers and big downloaders in bittorrent. In: LEET 2010, USENIX WORKSHOP ON LARGE-SCALE EXPLOITS AND EMERGENT THREATS. **Proceedings...** [S.l.: s.n.], 2010. p.4–4.

MANSILHA, R. B.; BAYS, L. R.; LEHMANN, M. B.; MEZZOMO, A.; GASPARY, L. P.; BARCELLOS, M. P. Observing the BitTorrent Universe Through Telescopes.

In: IM 2011, IFIP/IEEE INTERNATIONAL SYMPOSIUM ON INTEGRATED NETWORK MANAGEMENT. **Proceedings...** [S.l.: s.n.], 2011. p.1–8.

PIRATEBAY. Available at `http://thepiratebay.org`, access in July 2011.

PLANETLAB. Available at `http://www.planet-lab.org`, access in July 2011.

SCHULZE, H.; MOCHALSKI, K. **Internet Study 2008-2009**. Available at `http://www.ipoque.com/sites/default/files/mediafiles/documents/internet-study-2008-2009.pdf`, access in July 2011.

WIKIPEDIA. **List of Warez Groups**. Available at `http://en.wikipedia.org/wiki/List_of_warez_groups`, access in July 2011.

WIKIPEDIA. **Seedbox**. Available at `http://en.wikipedia.org/wiki/Seedbox`, access in July 2011.

ZHANG, C.; DHUNGEL, P.; WU, D.; LIU, Z.; ROSS, K. BitTorrent Darknets. In: INFOCOM 2010, IEEE INTERNATIONAL CONFERENCE ON COMPUTER COMMUNICATIONS. **Proceedings...** [S.l.: s.n.], 2010. p.1460–1468.

ZHANG, C.; DHUNGEL, P.; WU, D.; ROSS, K. Unraveling the BitTorrent Ecosystem. **IEEE Transactions on Parallel and Distributed Systems**, [S.l.], v.22, n.7, p.1164–1177, 2010.

# APPENDIX A – PUBLISHED PAPER AT SBSEG 2011

In this appendix, the paper entitled “Rumo à Caracterização de Disseminação Ilegal de Filmes em Redes BitTorrent” is presented. This was the first publication on this dissertation topic in a renowned scientific conference. Behavioral characteristics of those producing, publishing and providing illegal copies of movies are presented and analyzed.

- **Title:**
  *Rumo à Caracterização de Disseminação Ilegal de Filmes em Redes BitTorrent*
- **Conference:**
  11th Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBSeg 2011)
- **URL:**
  `http://www.ppgee.unb.br/sbseg2011/`
- **Data:**
  06-11 November of 2011
- **Location:**
  National University of Brasilia, Brasilia, Brazil

# Rumo à Caracterização de Disseminação Ilegal de Filmes em Redes BitTorrent

**Adler Hoff Schmidt, Marinho Pilla Barcellos, Luciano Paschoal Gaspary**

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{ahschmidt,marinho,paschoal}@inf.ufrgs.br

***Abstract.*** *Content sharing through BitTorrent (BT) networks accounts nowadays for a considerable fraction of the Internet traffic. Recent monitoring reports revealed that the contents being shared are mostly illegal and that movie is the most popular media type. Research efforts carried out to understand content production and sharing dynamics in BT networks do not provide precise information in respect to the behavior behind illegal film dissemination, being this the main objective and contribution of this paper. To perform such analysis, we monitored during 30 days all film torrent files published on the main BT public community. Furthermore, we joined the respective swarms, without downloading content, in order to obtain additional information regarding illegal sharing. As result, we present, characterize and discuss who produces and who publishes torrents of copyright-infringing files, what is produced and who acts as first provider of the contents.*

***Resumo.*** *O compartilhamento de conteúdo por meio de redes BitTorrent (BT) é atualmente um dos principais responsáveis pelo volume de dados na Internet. Relatórios de monitoração recentes constataram que os conteúdos sendo compartilhados são, em ampla maioria, ilegais e que filme é o tipo de mídia mais comum. Esforços de pesquisa realizados para entender a dinâmica de produção e de compartilhamento de conteúdo em redes BT não oferecem informações precisas sobre o comportamento por trás da disseminação ilegal de filmes, sendo esse o principal objetivo e contribuição deste artigo. Para realizar tal análise, monitorou-se todos os arquivos torrent de filmes publicados na principal comunidade pública de BT durante 30 dias e ingressou-se nos enxames, sem compartilhar conteúdo, a fim de obter informações adicionais acerca de compartilhamento. Como resultado, apresenta-se, caracteriza-se e discute-se quem produz e quem publica torrents de cópias ilícitas, o que é produzido e quem atua como primeiro provedor dos conteúdos.*

## 1. Introdução

Redes BitTorrent (BT) são atualmente a principal opção para usuários compartilharem conteúdo através da Internet [Schulze and Mochalski 2009]. Segundo estudo apresentado pela Envisional [Envisional 2011], aproximadamente dois terços dos 2,72 milhões de *torrents* administrados pelo principal rastreador BT são de conteúdos ilícitos, algo que reforça a noção intuitiva de BT ser largamente utilizado para compartilhar arquivos que infringem direitos autorais. O mesmo estudo aponta que 35,2% desses conteúdos ilícitos são cópias ilegais de filmes.

Apesar de existirem algumas publicações caracterizando compartilhamento de conteúdo em redes BT [Zhang et al. 2010b, Zhang et al. 2010a, Le Blond et al. 2010, Cuevas et al. 2010], nenhuma concentrou-se em observar aspectos específicos do processo de disseminação de conteúdos ilícitos, e muito menos de cópias ilegais de filmes (como, por exemplo, usuários responsáveis pela criação das cópias, tecnologias de digitalização utilizadas, etc). Pouco se sabe, por exemplo, sobre quem são os usuários responsáveis por criar cópias ilegais, quais são os processos de digitalização empregados, quem publica *torrents* desses conteúdos, e quem fomenta, nos estágios iniciais, os enxames formados em torno de cópias ilegais.

Algumas razões que justificam a importância de entender o compartilhamento ilegal de filmes em redes BT são discutidas a seguir. Primeiro, esse tipo de conteúdo é o principal responsável pelo volume de tráfego dessas redes. Segundo, caracterizar fidedignamente a atividade dos disseminadores desses conteúdos é base para formular mecanismos de combate a esse comportamento indesejado. Terceiro, responsáveis por conteúdos (no caso deste artigo, filmes) protegidos por direitos autorais podem amparar-se em conhecimento acerca do comportamento de usuários mal intencionados e criar estratégias para minimizar proliferação indevida de cópias ilegais.

Diante do problema e da motivação em abordá-lo, neste artigo apresenta-se resultados de um estudo experimental sistemático realizado para caracterizar disseminação ilegal de filmes em redes BT. Procura-se desvendar quem produz e quem publica cópias ilícitas, o que é produzido e quem atua como primeiro provedor. Além disso, estabelece-se relações entre agentes envolvidos e realiza-se exercício visando observar dinâmicas existentes (e não facilmente perceptíveis) no processo de disseminação ilegal de filmes. Para realizar o estudo, estendeu-se e utilizou-se a arquitetura de monitoração TorrentU [Mansilha et al. 2011], desenvolvida pelo grupo. Em um mês de monitoração, obteve-se 11.959 *torrents*, 1.985 nomes de usuários da comunidade, 94 rastreadores e 76.219 endereços IP únicos. Observou-se, ainda, atividades realizadas por 342 grupos de digitalização.

O restante do artigo está organizado como segue. A seção 2 apresenta conceitos acerca de compartilhamento ilícito de filmes e discute trabalhos relacionados. A seção 3 discorre sobre a arquitetura de monitoração e as decisões tomadas quanto à sua instanciação. A seção 4 relata e discute os resultados obtidos. A seção 5 encerra o artigo com considerações finais e perspectivas para trabalhos futuros.

## 2. Fundamentos e Trabalhos Relacionados

Esta seção está organizada em duas partes. Primeiro, revisita práticas adotadas por grupos de digitalização e características dos processos utilizados para digitalizar conteúdo. Na sequência, descreve e discute os trabalhos relacionados de maior relevância.

### 2.1. Conceitos Associados ao Compartilhamento Ilícito de Filmes

Grupos de digitalização são os responsáveis pela criação, através de meio ilícitos, de cópias de filmes [Wikipedia 2011a]. Eles podem ser compostos por um ou mais membros e recebem crédito pela sua atividade agregando a sua identificação ao nome dos *torrents* por eles criados. Via de regra, consumidores experientes não reconhecem um *torrent* de filme como confiável (e evitam usá-lo) caso ele não identifique o grupo de digitalização responsável. Logo, essa "etiqueta" é obedecida tanto por produtores quanto por consumidores. Ela provê uma maneira de dar notoriedade aqueles que estão realizando essa atividade ilegal, ao mesmo tempo em que os grupos buscam, para preservar sua reputação, assegurar o "casamento" correto entre nome dos *torrents* e conteúdos, bem como a correta classificação da qualidade desses *torrents*.

A decisão de usuários em realizar (ou não) *download* de determinadas cópias é influenciada, também, pela identificação, nos *torrents*, dos processos de digitalização empregados [Wikipedia 2011b]. A tabela 1 lista oito processos amplamente utilizados. Cada um deles é caracterizado por uma sigla, por uma fonte, *i.e.*, a mídia a partir da qual a cópia ilícita é gerada, e por uma expectativa de tempo, após a primeira estréia oficial do filme, para se encontrar uma cópia autêntica gerada usando o processo em questão. Os processos aparecem na tabela em ordem crescente de qualidade resultante esperada para as cópias digitalizadas.

**Tabela 1. Processos de digitalização**

| Sigla | Fonte | Lançamento |
|---|---|---|
| CAM | Gravado no cinema | 1 Semana (S) |
| TS | Gravado no cinema com fonte exclusiva de áudio | 1 S |
| TC | Material sendo projetado no cinema | 1 S |
| PPVRip | Exibição para clientes de hotéis | 8 S |
| SCR | Cópia distribuída a críticos e usuários especiais | Imprevisível |
| DVDScr | DVD distribuído para usuários especiais | 8-10 S |
| R5 | DVD não editado, lançado somente na região 5 | 4-8 S |
| DVDRip* | DVD acessível ao público | 10-14 S |

* Digitalizações a partir de fontes de maior qualidade, como Blu-ray, foram consideradas DVDRip.

## 2.2. Trabalhos Relacionados

Nos últimos anos a comunidade de pesquisa em redes par-a-par produziu alguns trabalhos ligados à monitoração de redes BT. Nesta seção descreve-se e discute-se os mais relacionados ao presente artigo. Por questão de escopo, são organizados em três grupos: infraestruturas de monitoração, caracterizações gerais do "universo" BT e caracterizações detalhadas de produção e consumo em redes BT.

Bauer *et al.* [Bauer et al. 2009] propuseram uma infraestrutura de monitoração que realiza medições ativas. A monitoração consiste em contatar rastreadores, obter endereços IP e contatar *hosts*, para confirmá-los como participantes "válidos" de enxames BT. Jünemann *et al.* [Junemann et al. 2010] desenvolveram uma ferramenta para monitorar *Distributed Hash Tables* (DHT) associadas a enxames BT. A ferramenta divide-se em três módulos. O primeiro permite coletar dados da rede par-a-par, como a quantidade de pares, endereços IP, portas utilizadas e países de origens, ao percorrer a DHT. O segundo analisa os dados e gera gráficos de acordo com métricas definidas pelos usuários. O terceiro busca, nos resultados do segundo módulo, valores que excedam limiares estipulados pelo usuário, gerando avisos. Ainda no campo de infraestruturas de monitoração, Chow *et al.* [Chow et al. 2007] apresentaram BTM: um sistema para auxiliar a detecção de pirataria que lança mão de monitoração automática de enxames BT. Ele é organizado em módulos responsáveis, respectivamente, pela procura de *torrents* na rede e pela análise dos mesmos. O discernimento entre quais dos materiais monitorados violam direitos autorais, e quais não, é completamente realizado pelo usuário através das regras que podem ser definidas para processamento dos dados coletados.

No que se refere a caracterizações gerais do "universo" BitTorrent, Zhang *et al.* [Zhang et al. 2010b] analisaram *torrents* de cinco comunidades públicas. A descoberta de pares deu-se através de comunicação com rastreadores ou consulta a DHTs. Os autores apresentam, entre outros aspectos, quais são as principais comunidades de BT, os graus de participação de cada publicador de *torrents*, as cargas e localizações dos principais rastreadores, a distribuição geográfica dos pares e as implementações de clientes BT mais utilizadas. Seguindo uma metodologia similar à desse trabalho, Zhang *et al.* [Zhang et al. 2010a] realizaram uma investigação sobre *darknets* em BT, *i.e.*, comunidades privadas. Entre os resultados apresentados, os autores comparam características de enxames impulsionados por *darknets* com de enxames "oriundos" de comunidades públicas. Como observação geral desses dois estudos, ressalta-se o interesse em "fotografar" momentos do ciclo de vida de enxames BT na tentativa de quantificá-los e de abstrair modelos. Não fez parte de seu escopo, contudo, analisar dinâmica e caracterizar padrões de disseminação de conteúdo ilícito.

Passando ao último grupo de trabalhos analisados, Blond *et al.* [Le Blond et al. 2010] monitoraram por 103 dias as três comunidades de BT mais populares, traçando perfis dos provedores de conteúdo e dos consumidores mais participativos. Conseguiram identificar 70% dos provedores, listar os principais conteúdos sendo compartilhados e caracterizar os participantes mais ativos (pares presentes em vários enxames). Cuevas *et al.* [Cuevas et al. 2010] investigaram os fatores socioeconômicos de redes BT, ressaltando os incentivos que os provedores de conteúdos têm para realizar essa atividade. Três grupos de publicadores foram definidos: os motivados por incentivos financeiros, os responsáveis por material falso e os altruístas. Com um mês de medições, esses grupos foram caracterizados por *Internet Service Providers* (ISPs) aos quais estão associados, tipos de conteúdos disponibilizados, incentivos para as suas atividades e renda monetária especulada. Os trabalhos de Blond *et al.* e Cuevas *et al.* são os que mais se assemelham ao apresentado neste artigo, em especial no nível detalhado de monitoração e nas técnicas empregadas. Destaca-se, entretanto, que o escopo desses trabalhos não foi o de analisar disseminação ilegal de filmes nem tampouco de conteúdo ilícito em geral. Logo, aspectos que desempenham papel importante na compreensão de esquemas ilegais de distribuição foram deixados de lado. Além disso, os trabalhos parecem apresentar limitações técnicas importantes. A título de exemplo, incluem nas estatísticas resultados associados a *torrents* com pouco tempo de vida nas comunidades (forte indicador de conteúdo falso), comprometendo análises realizadas e conclusões obtidas.

Nesta subseção revisou-se alguns dos trabalhos mais relevantes e correlatos a este artigo. Observa-se um esforço da comunidade de pesquisa em redes par-a-par em criar ferramental de monitoração e conduzir caracterizações. Nenhum dos trabalhos, porém, preocupou-se em investigar como redes BT vêm sendo usadas para disseminação ilegal de filmes e de outros conteúdos ilícitos. Acredita-se ser este um tópico de grande relevância, em especial para que se possa, com conhecimento de dinâmicas até agora obscuras, subsidiar a proposição de estratégias e mecanismos eficazes que propiciem a proteção de conteúdo protegido por direito autoral e, até mesmo, contribuir para ampliação do uso de redes BT em cenários mais sensíveis. Até onde sabemos, este é o primeiro trabalho que procura mapear, de forma sistemática, processo de disseminação de conteúdo ilegal em redes BT. As próximas seções detalham a arquitetura de monitoração empregada, aspectos de sua instanciação e os principais resultados obtidos.

## 3. Infraestrutura de Monitoração Utilizada

Esta seção apresenta a infraestrutura de monitoração utilizada. A subseção 3.1 apresenta a arquitetura de monitoração empregada, denominada TorrentU, e extensões implementadas para permitir a caracterização almejada. Em seguida, a subseção 3.2 detalha como a arquitetura foi instanciada.

### 3.1. Arquitetura de Monitoração TorrentU e Extensões

TorrentU [Mansilha et al. 2011] é uma arquitetura flexível projetada e desenvolvida para permitir a monitoração de redes BitTorrent. Como a figura 1 ilustra, a arquitetura segue a abordagem clássica gerente/agente e, portanto, possui basicamente dois componentes: observador e telescópios. Observador é o componente que faz o papel de *front-end*, isto é, gerente, permitindo que o operador configure o sistema e observe os dados coletados em tempo real (assim como o histórico dos dados). Telescópios, por sua vez, atuam como agentes, sendo os componentes responsáveis pela monitoração do universo BitTorrent e pelo retorno de resultados de acordo com as requisições enviadas pelo Observador. Telescópios são subdivididos em três partes, denominadas "lentes", sendo cada uma responsável por monitorar um grupo diferente de elementos do universo: comunidades, rastreadores e pares. Essa modularização permite que as lentes existentes possam ser substituídas, assim como novas possam ser facilmente incorporadas na arquitetura (sem modificação de seus componentes essenciais).

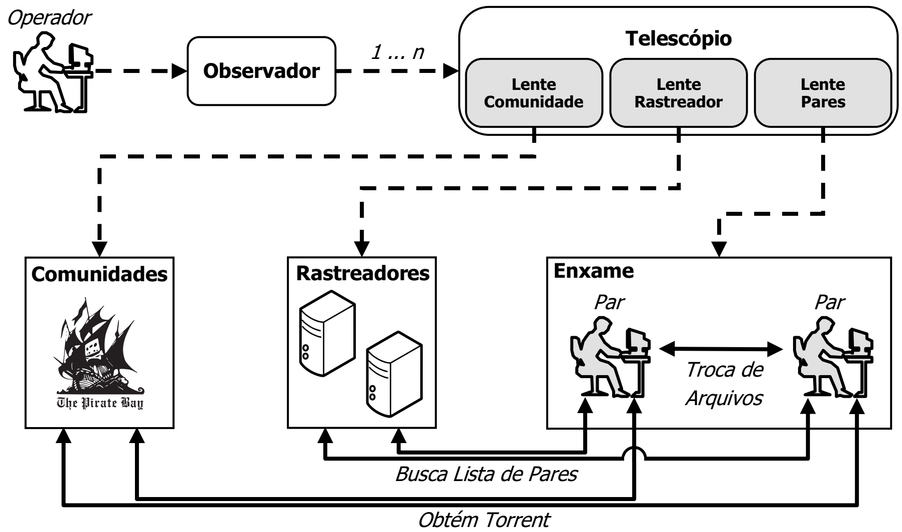



**Figura 1. Arquitetura TorrentU**

Lançando mão da flexibilidade oferecida por TorrentU, algumas funcionalidades, originalmente não contempladas pela arquitetura, foram implementadas e integradas. Entre as extensões, destacam-se duas. A primeira, criada para permitir a identificação dos primeiros pares semeadores de enxames, consiste em lente que captura *torrents* logo que publicados em comunidades. A segunda extensão, também materializada por meio de nova lente, realiza monitoração contínua das páginas de *torrents*, armazenando tempo de vida dos enxames, números de semeadores e sugadores, e testemunhos postados. O objetivo, nesse caso, é a produção de "fotografias" de enxames ao longo do tempo.

O algoritmo 1 apresenta uma visão geral do procedimento de monitoração executado. Como pode-se perceber pela descrição das funções, os *torrents* recentemente publicados são capturados. Para cada *torrent*, o(s) respectivo(s) rastreador(es) é/são caracterizado(s) e mensagem(ens) para obtenção de lista(s) de pares participantes, enviada(s). Caso obtenha-se essa(s) lista(s), um processo de caracterização dos primeiros pares participantes do enxame é realizado e, ao finalizá-lo, a lente responsável pela captura de fotografias da comunidade é iniciada para o *torrent* em questão. São cinco parâmetros que determinam o comportamento desse algoritmo. *Tempo* determina quanto durará a campanha de monitoração. *Rodadas* indica o número de tentativas a serem realizadas para contatar rastreadores. *Quantidade* representa o tamanho da lista de pares requisitada aos rastreadores. *Limiar* determina em quais enxames será feita troca de mensagens *bitfield*. *Periodicidade* consiste no intervalo de tempo a ser respeitado para produzir cada fotografia de um dado enxame. *Intervalo* representa o tempo de espera entre cada rodada de execução do algoritmo.

```
input: tempo, tentativas, quantidade, limiar, periodicidade e intervalo
for i ← 0 to tempo do
    torrents[] ← CapturarTorrentsRecentes();
    for j ← 0 to torrents.size() do
        torrent ← torrents[j];
        DownloadTorrent(torrent);
        LerArquivo(torrent);
        CaracterizarRastreadores(torrent);
        peerList ← ObterListaPares(torrent, tentativas,
        quantidade);
        CaracterizarPares(peerList);
        if peerList.size() < limiar then
            TrocarBitfields(torrent);
        end
        IniciarCapturaFotografias(torrent, periodicidade);
    end
    Esperar(intervalo);
end
```

**Algoritmo 1:** Visão geral do procedimento de monitoração

Ressalta-se que a ênfase deste artigo reside nos resultados da caracterização de disseminação ilegal de filmes e não na descrição da arquitetura de monitoração. Ao leitor interessado em detalhes acerca do funcionamento da arquitetura sugere-se consulta a artigo anterior [Mansilha et al. 2011] produzido pelo nosso grupo de pesquisa.

### 3.2. Instanciação da Arquitetura

Telescópios foram instanciados em três nodos do PlanetLab [PlanetLab 2011] e em um servidor privado. O objetivo dessa redundância foi, basicamente, tolerar falhas e evitar descontinuidade do processo de monitoração. Já o componente Observador foi instanciado em uma única estação. Entre as comunidades BitTorrent existentes, optou-se por monitorar o PirateBay [Piratebay 2011]. Tal deve-se ao fato de ser a comunidade aberta mais popular, disponibilizar somente *torrents* publicados em seus servidores, manter registro de usuários responsáveis pela publicação de cada *torrent*, e prover classificação de cada usuário baseada em sua reputação.

O processo de monitoração foi instanciado utilizando-se as seguintes configurações (podem ser entendidas como parâmetros recém detalhados do algoritmo 1): 2 tentativas para obtenção de lista de pares com cada rastreador, 50 pares por lista, limiar definindo máximo de 10 pares com os quais serão trocadas as mensagens *bitfield*, periodicidade de 8 horas entre cada fotografia da comunidade e intervalos de 2 minutos entre rodadas de monitoração. A monitoração foi realizada por período de um mês (05/2011 à 06/2011), produzindo dados brutos que somaram 11.959 *otorrents*, 94 rastreadores e 187.140 endereços IP.

## 4. Resultados

A base de 11.959 *torrents* precisou ser submetida a um processo de filtragem, para que enxames indesejados fossem retirados e, assim, não influenciassem a análise. Inicialmente removeu-se todos os *torrents* cujos rastreadores não puderam ser contatados devido a inconsistências nas URLs informadas.. Nesse grupo enquadraram-se 4.181 *torrents*. Em um segundo momento, retirou-se aqueles *torrents* que tiveram menos de 8 horas de vida na comunidade, levando à glosagem de mais 4.791 *torrents*. Enxames com dados inconsistentes ou removidos tão precocemente da comunidade representam, muito provavelmente, conteúdos falsos. A não remoção desses enxames pode exercer forte influência na análise dos resultados. Apesar da importância do processo de filtragem, até onde sabemos em nenhum dos trabalhos relacionados houve tal preocupação.

Após as duas filtragens, 2.987 *torrents* remanesceram e foram analisados. Os resultados são apresentados a seguir. Inicialmente caracteriza-se produtores e publicadores de conteúdos ilícitos, investigando seus graus de atividade e possíveis relações entre esses agentes. Na sequência, relata-se os processos de digitalização mais empregados e detalha-se a dinâmica, ao longo do tempo, de lançamento de *torrents* (dado um conjunto conhecido de conteúdos) *x* processos utilizados. Por fim, analisa-se os primeiros semeadores, procurando relações entre eles, os criadores de cópias digitais e os publicadores de *torrents*.

### 4.1. Produtores e Publicadores de Conteúdo Ilícito

Conforme apresentado na subseção 2.1, grupos de digitalização são os principais responsáveis pela criação de cópias ilícitas de filmes sendo compartilhadas nas redes BT. Neste artigo eles são referidos como *produtores*. Dos 2.987 *torrents* analisados, 2.066 (69,16%) identificam o produtor responsável por cada cópia. Esses 2.066 *torrents* foram criados por 342 produtores distintos. A tabela 2 enumera, em ordem decrescente, os 10 principais produtores. Ao lado, na figura 2, ilustra-se CDF (*Cumulative Density*

*Function*) representando no eixo horizontal os produtores, ordenados por quantidade de conteúdo criado, e no eixo vertical, a proporção acumulada de *torrents* criados. Como é possível observar, poucos produtores são responsáveis por grande parcela do conteúdo criado; praticamente 80% das cópias foram criadas por 100 produtores (29,23% dos 342).

**Tabela 2. Ranqueamento**

| Grupo | Torrents # | Torrents % |
|---|---|---|
| Waf | 78 | 4,02 |
| Mr_Keff | 69 | 3,56 |
| Tnt Village | 62 | 3,20 |
| DutchTeamRls | 61 | 3,14 |
| Dmt | 61 | 3,14 |
| Imagine | 59 | 3,04 |
| Lkrg | 51 | 2,63 |
| Miguel | 46 | 2,37 |
| Martin | 46 | 2,37 |
| Nlt | 44 | 2,27 |

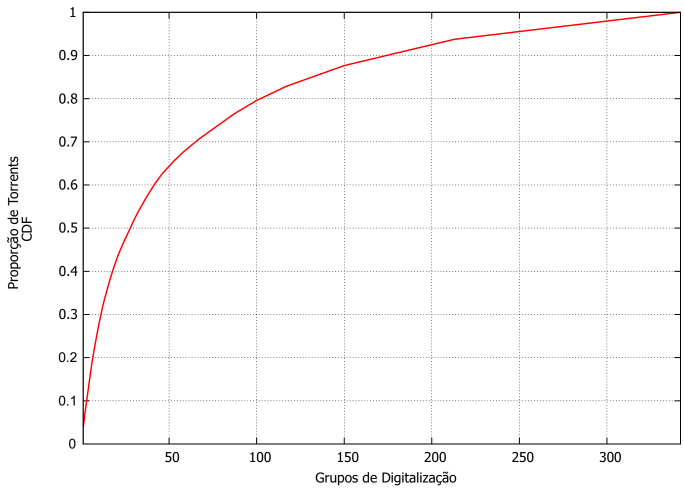



**Figura 2. Contribuição cumulativa dos *produtores***

Com o arquivo digital criado, o próximo passo para disseminá-lo é a publicação de *torrent* na comunidade. Os responsáveis por essa etapa são os publicadores, que, no escopo deste artigo, correspondem a usuários cadastrados no PirateBay realizando *upload* de *torrents*. Os 2.987 *torrents* foram publicados por um total de 517 usuários distintos. A tabela 3 apresenta os usuários mais ativos junto com a quantidade e proporção de *torrents* publicados. Essa tabela apresenta, além do nome do usuário, a sua categoria, que representa um "termômetro" da sua reputação na comunidade. Existem quatro categorias de usuários: VIP, confiável, ajudante e regular (estado inicial de qualquer usuário). A figura 3 apresenta uma CDF com os usuários em ordem decrescente de *torrents* publicados no eixo horizontal e a proporção de *torrents* no vertical. Ao observar esse gráfico, pode-se, novamente, perceber como poucos usuários são responsáveis pela maioria do conteúdo publicado. Em números, tem-se que 100 usuários (19,34% dos 517) publicaram quase 75% do conteúdo. Além disso, destaca-se que 25,5% dos 517 usuários eram de tipos especiais (não regulares) e foram eles os responsáveis pela publicação de 59,9% dos *torrents*.

Após análise isolada da atividade de produtores e consumidores, procurou-se evidências quanto à existência de relação na ação de ambos agentes. Dois casos típicos foram observados: um publicador disponibilizando todos os materiais de um produtor e um grupo de publicadores trabalhando para um único produtor. Exemplificando o primeiro caso tem-se o usuário "sadbawang", que publicou 77 dos 78 *torrents* do grupo "Waf". Para ilustrar o segundo caso, tem-se que os publicadores ".BONE." e "froggie100" foram responsáveis pela maioria dos conteúdos criados pelo grupo "Imagine".

### 4.2. Processos de Digitalização Empregados

Como já apontado na subseção 2.1, cópias digitais de um mesmo filme são diferenciadas pelas suas qualidades, que, por sua vez, são resultantes do processo de digitalização uti-

**Tabela 3. Ranqueamento**

| Usuário | Tipo | Torrents # | Torrents % |
|---|---|---|---|
| .BONE. | VIP | 211 | 7,06 |
| HDVideos | Regular | 91 | 3,04 |
| sadbawang | VIP | 77 | 2,57 |
| Sir_TankaLot | Confiável | 73 | 2,44 |
| MeMar | VIP | 67 | 2,24 |
| l.diliberto | VIP | 57 | 1,90 |
| SaM | VIP | 47 | 1,57 |
| martin_edguy | Confiável | 46 | 1,54 |
| miguel1983 | VIP | 46 | 1,54 |
| virana | Confiável | 41 | 1,37 |

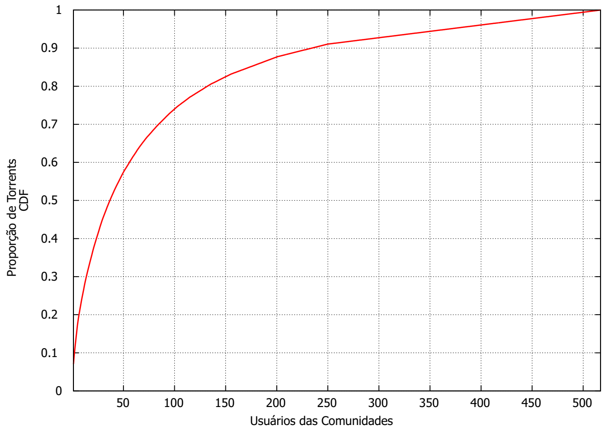



**Figura 3. Contribuição cumulativa dos *publicadores***

lizado. Dos 2.987 *torrents* analisados, 2.344 (78,47%) identificavam o processo utilizado para criação de cada mídia. A tabela 4 apresenta os processos, os seus graus de ocorrência e os três grupos de digitalização que mais criaram mídias empregando cada processo.

**Tabela 4. Principais processos**

| Processo | Torrents # | Torrents % | Principais Grupos |
|---|---|---|---|
| DVDRip | 1825 | 77,85 | Waf, Mr_Keff, Tnt Village |
| TS | 144 | 6,14 | Imagine, Dtrg, Cm8 |
| R5 | 132 | 5,63 | Imagine, Dmt, Vision |
| DVDScr | 109 | 4,65 | Ddr, Mtr, Xtreme |
| CAM | 68 | 2,90 | Lkrg, Imagine, Team Tnt |
| PPVRip | 27 | 1,15 | Iflix, Dmt, Flawl3ss |
| SCR | 22 | 0,93 | Kickass, Scr0n, 7speed |
| TC | 16 | 0,68 | Mtr, Team Tc, Rko |

Analisando os resultados, observa-se que “DVDRip” representa o processo de digitalização mais comum, sendo o empregado por 77,85% dos *torrents* analisados que identificaram o processo. Tal predominância deve-se a duas razões principais. Primeiro, o conjunto de filmes que esses *torrents* podem estar representando é muito maior. Qualquer filme com 16 semanas transcorridas do seu lançamento oficial pode ser encontrado nesse formato e o resultado desse processo são mídias com a qualidade máxima, resultando no desinteresse pelas criadas por outros processos. Segundo, digitalizar um filme por meio desse processo é trivial se comparado com os outros; qualquer usuário que possuir um DVD original pode fazê-lo em seu computador pessoal. Outro grupo de processos que se destaca é o formado por “CAM”, “TS”, “DVDScr” e “R5”. O alto grau de ocorrência, em comparação aos outros processos que não “DVDRip”, deve-se a uma questão de custo/benefício entre: a dificuldade de obter-se a fonte para digitalização, o tempo necessário após o lançamento oficial do filme e a qualidade final da mídia. A título de exemplo tem-se que, apesar da qualidade resultante do processo “TC” ser a melhor entre a estréia do filme e 4-8 semanas transcorridas, “CAM” e “TS”, por utilizarem fonte facilmente acessíveis, são processos mais empregados (0,68% *x* 9,04%). Vale observar, também, como produtores "menos ativos" destacam-se pelas suas especializações em processos que necessitam de fontes de difícil acesso. O grupo "Imagine", por exemplo, apesar de ser somente o sexto colocado da tabela 2, apresenta-se como o principal produtor de "TS" e "R5". Logo, as atividades desse grupo tornam-se tão importantes quanto, se não mais, as dos primeiros colocados da tabela 2.

Passando-se, agora, a caracterizar a dinâmica de processos de digitalização e de *torrents* disponibilizados em portais BT em paralelo ao ciclo de vida de filmes lançados no cinema, a figura 4 sintetiza resultado de observação realizada. Antes de apresentar discussão, contudo, é necessário informar que o período mínimo de monitoração para ser possível capturar todas as digitalizações realizadas sob um mesmo filme é de 16 semanas. Como não dispôs-se desta janela de tempo, trabalhou-se com 9 filmes, todos presentes no *dataset* coletado ao longo de 30 dias, cujo lançamento tivesse ocorrido há 0-4, 5-8 e há mais de 8 semanas (de acordo com o informado na IMDb [IMDB 2011]). No gráfico, o eixo horizontal representa os dias transcorridos e o eixo vertical, os processos de digitalização. Para apresentar uma "fotografia" mais fidedigna, todos *torrents* que não tiveram um tempo mínimo de vida de uma semana na comunidade e que não haviam sido publicados por usuários renomados foram desconsiderados.

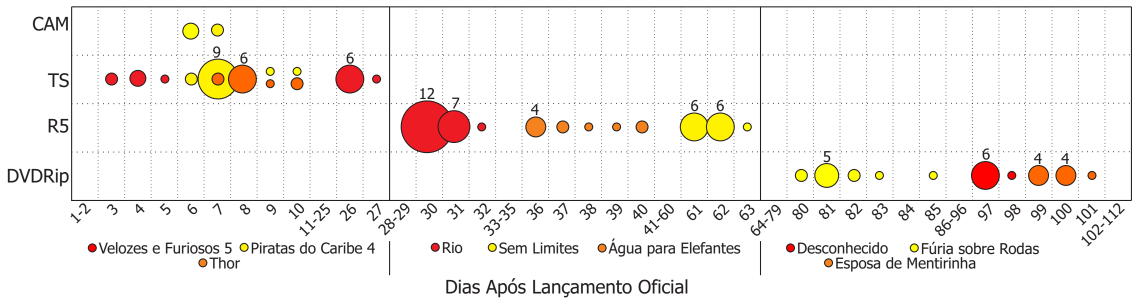



**Figura 4. Processos de digitalização utilizados após estréia oficial**

Quatro aspectos merecem destaque a partir da análise do gráfico. Primeiro, mídias criadas por um processo surgem em rajadas de *torrents*, cada um representando o trabalho de um grupo distinto. Tal pode ser observado, por exemplo, nos dias 30 e 31, em que surge a primeira mídia gerada pelo processo "R5" do filme "Rio". Segundo, como esperado, os intervalos apresentados na tabela 1 para surgimento das mídias geradas por meio de cada processo são respeitados. O momento exato dentro desse intervalo é influenciado por decisões da indústria cinematográfica. Por exemplo, os primeiros arquivos gerados pelo processo "R5" dos filmes "Rio", "Água para Elefantes" e "Sem Limites" aparecem, respectivamente, quatro, cinco e oito semanas após as suas estréias, pois as suas fontes foram lançadas em momentos distintos. Terceiro, *torrents* de alguns filmes continuam sendo publicados mesmo após o final da rajada inicial, como ocorre nos dias 82, 83 e 85 do filme "Fúria sobre Rodas". Tal não deve, contudo, ser encarado como indício de comportamento de distribuição diferenciado; trata-se de especializações de mídias já existentes (codificação de vídeo alternativa, áudio dublado ou legenda inserida sobre a imagem do vídeo). Quarto, o mesmo filme pode ter mais do que uma rajada de publicações por processo de digitalização. Observa-se essa situação analisando o filme "Velozes e Furiosos 5", em que uma rajada inicia-se no dia 3 e outra no dia 26. Esse fenômeno é observado quando uma fonte de melhor qualidade é encontrada para realizar o processo, acarretando em melhor qualidade final da mídia gerada.

## 4.3. Provedores de Conteúdo Ilícito

Ainda como parte da caracterização de disseminação ilegal de filmes em redes BT, interessou-se em determinar os pares (usuários) que fomentam enxames, na condição de semeadores, em seus instantes iniciais. Para identificá-los, foi necessário que a arquitetura de monitoração estivesse devidamente configurada para, assim que *torrents* fossem publicados no portal, pudessem ter seus respectivos rastreadores contatados e listas de pares participantes do início do enxame, obtidos. Quanto menor o intervalo decorrido entre a publicação de um *torrent* e o início da monitoração do enxame, maior a probabilidade de encontrar no enxame apenas o(s) primeiro(s) semeador(es). No contexto da investigação conduzida (lançando mão da arquitetura TorrentU e, em última análise, do procedimento ilustrado pelo algoritmo 1), esse intervalo girou em torno de 4 minutos.

Por meio da metodologia mencionada, identificou-se os primeiros semeadores de 692 (23,16%) dos 2.987 *torrents* analisados. Todos aqueles em que não foi possível identificar o(s) primeiro(s) semeador(es) eram enxames que: estavam vazios, existiam previamente à publicação do seu *torrent* no PirateBay ou cujo(s) rastreador(es) não foi possível contatar por erro na tentativa de comunicação. Passando à análise dos resultados, os 692 *torrents* foram semeados por um total de 775 pares; para alguns enxames observou-se, já no seu início, mais do que um semeador. Os 775 semeadores identificados estão associados a 318 endereços IP únicos. A figura 5 ilustra o grau de participação de cada IP.

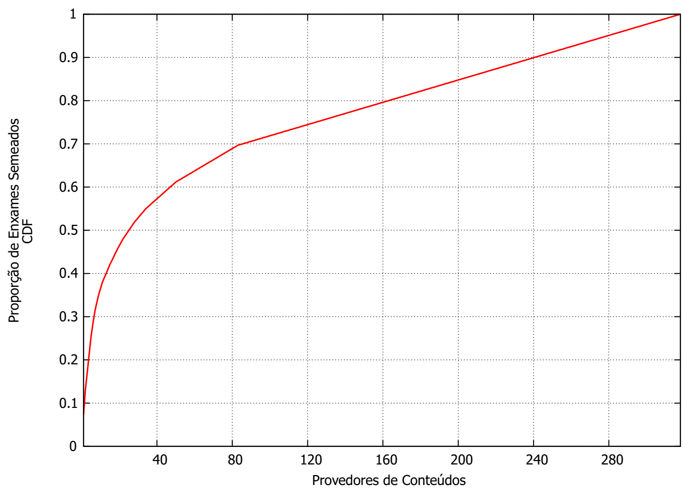



**Figura 5. Contribuição cumulativa dos *provedores***

Dois aspectos destacam-se a partir da análise da figura 5. Primeiro, 25 endereços IP (7,86% dos 318) participaram como semeadores de cerca da metade dos enxames. Tal é um indicador de que usuários especializados podem estar utilizando *seedboxes* [Wikipedia 2011c] para disseminar seus conteúdos. Segundo, todos os semeadores a partir do 82º serviram exclusivamente a um enxame, caracterizando a participação de usuários "domésticos" no fomento de parcela significativa de enxames BT.

A tabela 5 apresenta a procedência dos principais semeadores, destacando país de origem, *Internet Service Provider* (ISP) de cada IP e quantidade de enxames semeados. Em contraste, apresenta-se na tabela 6 as principais localizações dos semeadores, ignorando a quantidade de enxames que cada um serviu. Como pode-se observar, a França destaca-se por 9,03% dos semeadores estarem localizados nesse país, curiosamente sendo todos servidos pelo ISP "Ovh".

| País | ISP | # Enxames |
|---|---|---|
| NZ | Obtrix | 57 |
| FR | Ovh | 45 |
| FR | Ovh | 32 |
| PL | Mokadi | 32 |
| GB | Ovh | 31 |
| FR | Ovh | 24 |
| NZ | Obtrix | 21 |
| FI | Lsinki | 15 |
| FR | Ovh | 14 |
| NL | Upc | 12 |

**Tabela 5. Principais semeadores**

| País | # IPs |
|---|---|
| IN | 41 |
| US | 33 |
| SE | 33 |
| FR | 28 |
| NL | 26 |
| JP | 24 |
| GB | 14 |
| PK | 11 |
| DE | 10 |
| AU | 7 |

**Tabela 6. Distrubuição semeadores**

Os provedores de conteúdo são os terceiros e últimos agentes responsáveis pelo processo de disseminação de cópias ilícitas de filmes através de BT. Ao observá-los, constatou-se que existem relações de dependência, ou subordinação, entre provedores (primeiros semeadores), produtores (grupos de digitalização) e publicadores (usuários da comunidade). Três casos típicos foram encontrados e são discutidos a seguir. O primeiro caso são provedores e publicadores dependentes dos produtores. Um exemplo são os grupos "Dmt", "Mr_keff" e "Miguel", que tiveram cerca de 90% dos seus *torrents* publicados e semeados pelo mesmo usuário. O segundo caso consiste na observação de que provedores podem estar subordinados a produtores. Exemplos desse caso são os grupos "Kickass", "Ddr" e "Extratorrentrg" que, apesar de terem seus *torrents* publicados por um grupo heterogêneo de usuários, sempre são semeados pelo mesmo provedor. O terceiro e último caso está relacionado com a possibilidade de provedores serem dependentes de publicadores. Como exemplo, tem-se os publicadores "Theroach", "Riff" e "Safcuk009", que disponibilizaram *torrents* de mídias criadas por grupos variados, porém sempre semeados pelos mesmos provedores.

## 5. Conclusões e Trabalhos Futuros

O compartilhamento de conteúdo por meio do protocolo BitTorrent é um dos principais responsáveis pelo atual volume de tráfego da Internet. Sabe-se que a maior parte desse volume é constituída pelo compartilhamento de conteúdos ilícitos. Sabe-se, também, que filme é o principal tipo de mídia sendo compartilhado ilegalmente. Apesar da reconhecida importância do tema, nenhum estudo procurou observar e mapear a dinâmica de disseminação dessa natureza de conteúdo. Para suprir essa lacuna, realizou-se a extensão de uma arquitetura de monitoração, que foi instanciada para observar o "universo" BT por 30 dias. A grande massa de dados obtida foi, então, organizada e cuidadosamente analisada. Até onde sabemos, este é o primeiro estudo científico que busca caracterizar disseminação ilegal de filmes em redes BitTorrent.

A partir dos dados obtidos foi possível identificar padrões de comportamento de disseminadores de filmes ilegais. No que remete aos produtores, descobriu-se que a maioria dos *torrents* possui identificação do grupo de digitalização responsável e que, na realidade, são poucos produtores criando a maioria dos conteúdos. Quanto aos publicadores, observou-se um comportamento similar ao dos produtores, no sentido de que poucos são responsáveis pela publicação de grande parte dos *torrents* de cópias ilícitas. Além disso, uma relação de subordinação foi observada entre produtores e publicadores. Ao analisar os processos de digitalização empregados, descobriu-se que certos produtores são especializados em certos processos e confirmou-se a evolução, ao longo do tempo, dos processos por meio dos quais os filmes ofertados são digitalizados. Por fim, abordou-se os responsáveis por inicialmente semearem os enxames desses conteúdos, identificando as relações de dependência existentes entre produtores, publicadores e semeadores.

Finalizada uma primeira iteração para caracterizar disseminação ilegal de filmes em redes BT, identifica-se um conjunto de oportunidades de investigações futuras. A primeira consiste em realizar monitoração mais longa, desejavelmente com um mínimo de 16 semanas, que permita acompanhar o "ciclo de vida" completo de filmes em redes BT, desde seu lançamento no cinema até o momento em que passa a ser distribuído em DVD. A segunda oportunidade de trabalho futuro consiste em observar outras comunidades públicas populares. A terceira, monitorar as *darknets*, comunidades privadas de *torrents*, que, provavelmente, representam os locais onde, em tese, enxames em torno de cópias ilegais de filmes aparecem primeiro. Por fim, uma quarta oportunidade é a observação de padrões e dinâmicas de consumo desses conteúdos. Por exemplo, é interessante procurar observar se há concentração de consumidores de filmes ilegais em determinadas regiões e se há comportamentos claros de migração de consumidores entre enxames.

## Referências


Bauer, K., McCoy, D., Grunwald, D., and Sicker, D. (2009). Bitstalker: Accurately and efficiently monitoring bittorrent traffic. In *WIFS 2009, IEEE Workshop on Information Forensics and Security*, pages 181–185.

Chow, K., Cheng, K., Man, L., Lai, P., Hui, L., Chong, C., Pun, K., Tsang, W., Chan, H., and Yiu, S. (2007). Btm - an automated rule-based bt monitoring system for piracy detection. In *ICIMP 2007, International Conference on Internet Monitoring and Protection*, pages 1–6.

Cuevas, R., Kryczka, M., Cuevas, A., Kaune, S., Guerrero, C., and Rejaie, R. (2010). Is content publishing in bittorrent altruistic or profit-driven? In *Co-NEXT 2010, International Conference on Emerging Networking Experiments and Technologies*, pages 1–12.

Envisional (2011). An estimate of infringing use of the internet. `http://documents.envisional.com/docs/Envisional-Internet_Usage-Jan2011.pdf` [Acessado em 07/11].

IMDB (2011). `http://www.imdb.com/` [Acessado em 07/11].

Junemann, K., Andelfinger, P., Dinger, J., and Hartenstein, H. (2010). Bitmon: A tool for automated monitoring of the bittorrent dht. In *P2P 2010, IEEE International Conference on Peer-to-Peer Computing*, pages 1–2.

Le Blond, S., Legout, A., Lefessant, F., Dabbous, W., and Kaafar, M. A. (2010). Spying the world from your laptop: identifying and profiling content providers and big downloaders in bittorrent. In *LEET 2010, USENIX Workshop on Large-Scale Exploits and Emergent Threats*, pages 4–4.

Mansilha, R. B., Bays, L. R., Lehmann, M. B., Mezzomo, A., Gaspary, L. P., and Barcellos, M. P. (2011). Observing the bittorrent universe through telescopes. In *IM 2011, IFIP/IEEE International Symposium on Integrated Network Management*, pages 1–8.

Piratebay (2011). `http://thepiratebay.org/` [Acessado em 07/11].

PlanetLab (2011). `http://www.planet-lab.org/` [Acessado em 07/11].

Schulze, H. and Mochalski, K. (2009). Internet study 2008-2009. `http://www.ipoque.com/sites/default/files/mediafiles/documents/internet-study-2008-2009.pdf` [Acessado em 07/11].

Wikipedia (2011a). List of warez groups. `http://en.wikipedia.org/wiki/List_of_warez_groups` [Acessado em 07/11].

Wikipedia (2011b). Pirated movie release types. `http://en.wikipedia.org/wiki/Pirated_movie_release_types` [Acessado em 07/11].

Wikipedia (2011c). Seedbox. `http://en.wikipedia.org/wiki/Seedbox` [Acessado em 07/11].

Zhang, C., Dhungel, P., Wu, D., Liu, Z., and Ross, K. (2010a). Bittorrent darknets. In *INFOCOM 2010, IEEE International Conference on Computer Communications*, pages 1460–1468.

Zhang, C., Dhungel, P., Wu, D., and Ross, K. (2010b). Unraveling the bittorrent ecosystem. *IEEE Transactions on Parallel and Distributed Systems*, 22(7):1164–1177.

# APPENDIX B – PUBLISHED PAPER AT NOMS 2012

In this appendix, the paper entitled "Characterizing Dissemination of Illegal Copies of Content through Monitoring of BitTorrent Networks" is presented. This was the second publication on this dissertation topic. This paper characterizes in a more thoroughly manner the users responsible for the dissemination of illegal copies of content in comparison to the first paper. Examples of such can be cited as: the monitoring process gathered data over a greater period of time, novel aspects of the interrelation among types of users disseminating were presented and some attempts to model the observed behavior were performed.

- **Title:**
  *Characterizing Dissemination of Illegal Copies of Content through Monitoring of BitTorrent Networks*
- **Conference:**
  IEEE/IFIP Network Operations and Management Symposium (NOMS 12)
- **URL:**
  `http://www.ieee-noms.org/index.html`
- **Data:**
  16-20 April of 2012
- **Location:**
  Maui, Hawaii, United States of America

# Characterizing Dissemination of Illegal Copies of Content through Monitoring of BitTorrent Networks

Adler Hoff Schmidt, Rodolfo Stoffel Antunes, Marinho Pilla Barcellos, Luciano Paschoal Gaspary
Institute of Informatics – Federal University of Rio Grande do Sul (UFRGS)
P. O. Box 15.064 – 91.501-970 – Porto Alegre – RS – Brazil
{ahschmidt,rsantunes,marinho,paschoal}@inf.ufrgs.br

***Abstract*—BitTorrent networks are nowadays the most employed method of Peer-to-Peer (P2P) file sharing in the Internet. Recent monitoring reports reveal that content copies being shared are mostly illegal and movies are the most popular media type. Research efforts carried out to understand the dynamics of content production and sharing in BT networks have been unable to provide precise information regarding the dissemination of illegal copies. In this paper we perform an extensive experimental study in order to characterize the behavior of producers, publishers and providers of copyright-infringing files. The study is based on four months of traces obtained by monitoring swarms sharing movies via one of the most popular BT public communities. Traces were obtained with an extension of a BitTorrent "universe" observation architecture, which allowed the collection of a database with information about more than 40,000 torrents, 900 trackers and 1.3 million IPs. Our analysis not only shows that a small group of active users is responsible for the majority of disseminated illegal copies, as well as unravels existing relationships among these actors.**

## I. Introduction

The BitTorrent (BT) protocol is currently the most used option for content sharing over the Internet [1]. A recent study by Envisional [2] shows that illegal copies of copyrighted content can be found in more than two thirds of torrents registered at one of the most popular BT trackers. Such number reinforces the common sense that BitTorrent is extensively used for sharing of copyrighted files. The same study also indicates that over one third of the illegal copies are movies.

Several studies, such as [3], [4], [5], [6], have been carried out in recent past to characterize content sharing in BT networks. None of these, however, focused on issues specific to the dissemination process of illegal copies. For example, little is known about trends in the behavior of users who: (*i*) obtain access to the original content in order to create digital copies of it; or (*ii*) publish these illegal copies in BT networks. Further, it is unclear to which degree the users who create digital copies of copyrighted content are the same ones that make the corresponding copies available in BT communities.

Protection mechanisms are necessary to effectively mitigate the dissemination of illegal copies of copyrighted content through file sharing mechanisms. Owners of protected content, in turn, would be interested in developing strategies in order to minimize the possibilities that their property will be copied and published through illegal means. Such goals, however, require the development of a body of knowledge related to the processes employed in the creation and dissemination of illegal copies in file sharing communities.

This paper presents results of an experimental study conducted in order to characterize the dissemination of illegal copies of content through file sharing communities. We seek to identify trends in the behavior of users who generate such copies and also of those who publish them. Our study focused on communities that employ the BT protocol because it is responsible for most of the P2P file sharing traffic over the Internet. Our observations were conducted with traces collected through extensions developed for the TorrentU monitoring architecture [7]. Since movies encompass a large portion of the observed illegal copies, our study will be focused on this type of content. Our results, however, can be generalized for other types, such as music and software. With traces recording four months of activities, we obtained 40,993 torrents, 7,235 community usernames, 915 trackers and more than 1.3 million IP addresses. We also observed the activities of 482 content digitalization groups.

From the distributed systems operations and management point of view, this paper follows the track of previous investigations performed by this community on the area (*e.g.*, [8], [9], [10], [11]). We believe an important contribution of our paper to the field is the presentation of fresh and in-depth characterization results – obtained by means of a long term, large-scale monitoring campaign – of the dynamics behind the dissemination of illegal copies of copyrighted content in BT networks. The results are deemed meaningful to Internet and multimedia service providers, to the film industry, as well as to a community of researchers who investigate strategies and mechanisms to promote a more secure usage of swarm-based content sharing systems. Furthermore, although not the main focus of the paper, we do revisit and extend an architecture (proposed in the context of our group), which is tailored to perform active, application-layer protocol monitoring.

The remaining of this paper is organized as follows. Section II presents concepts related to the sharing of illegal copies of movies and discusses related work. Section III explains the monitoring architecture and how it was instantiated. Section IV discusses the collected data and provides insights about it. Finally, Section V presents conclusions and perspectives of future work.

## II. Background & Related Work

The first part of this section presents some empirical information about the processes adopted by digitalization groups in order to generate illegal copies of movies. Next we discuss other studies focused on the characterization of content distributed through BitTorrent networks.

### A. Background: Illegal Copies of Movies

Digitalization groups are responsible for creating copies of movies through illegal methods [12]. They are composed by one or more members and claim merit for their activity by

978-1-4673-0269-2/12/$31.00 

adding their pseudonym to the created torrent files. Empirical observations of BT communities indicate that expert users do not recognize a torrent file as trustworthy (and avoid using it) if it does not contain the digitalization group identification. The use of a pseudonym in torrents allows digitalization groups to build a reputation, and groups seem to compete with each other in this respect. Thus, this pseudonym is observed by both content producers and consumers. A digitalization group also seeks to preserve its reputation with two methods: (*i*) ensuring that its pseudonym is not present in copies created by other groups; and (*ii*) guaranteeing that the digitalization process result maintains an expected quality level.

The users decision about downloading (or not) a specific copy is also influenced by the type of digitalization process indicated in the content information. During six months, we observed the publication of new movie-related torrents in popular communities. Our observation leads to the types of digitalization processes presented in Table I. Each one is identified by: (*i*) an acronym; (*ii*) a source (*i.e.*, the media that serves as basis for creation of the illegal copy); and (*iii*) the minimum expected time an illegal copy based on such digitalization method can be found after the original premiere of the movie. Processes in Table I are ordered according to the expected quality of the created copy. Our observations regarding quality of image and sound are essentially empirical, but firmly supported by comments posted in blogs and community sites. The expected release dates are applicable to the communities we monitored, but might be different in other communities (*e.g.*, private ones). It should be noted that the "DVDRip" process may be performed with sources of higher quality, such as Blu-ray discs. Copies employing sources of higher quality than DVD discs, however, are also identified as created with the "DVDRip" process.

Table I
DIGITALIZATION PROCESSES

| Acronym | Source | Estimated Time |
|---|---|---|
| CAM | Recorded at a movie theater | Aprox. 1 Week |
| TS | Recorded at a movie theater with exclusive audio source | Aprox. 1 Week |
| TC | Directly copied from theaters media | Aprox. 1 Week |
| PPVRip | Content exhibited to hotels clients | Aprox. 8 Weeks |
| SCR | Copy distributed to critics and special users | Unpredictable |
| DVDScr | DVD distributed to special users | Aprox. 8 Weeks |
| R5 | Non-edited DVD, launched only on region 5 | Aprox. 4 Weeks |
| DVDRip | DVD distributed to general public | Aprox. 10 Weeks |

### B. Related Work

Studies related to ours are divided in two classes. We first present a summary of proposed monitoring infrastructures for BT networks. Next, we review studies that focus on the observation of the BT "universe" in order to identify its general characteristics and to model the creation and distribution of content.

Bauer *et al.* [13] proposed a monitoring infrastructure based on active measurement of BT swarms. The monitoring consists in contacting trackers to obtain IP addresses from peers and then verifying these in order to acknowledge them as valid BT peers. Jünemann *et al.* [14] developed a tool to monitor distributed hash tables (DHT) associated with BT swarms. This tool is composed of three modules. The first allows the collection of data from the P2P network such as the number of peers and IP addresses and ports through queries to the DHT overlay. The second module analyzes the data and generates graphs according to predefined metrics. The third and final module generates warnings for situations such as torrents with high number of connected peers. Another monitoring infrastructure, named BTM, is presented by Chow *et al.* [15]. It focuses on the detection of piracy through automatic monitoring of BT swarms. The BTM architecture is organized in two modules: one for searching torrent files and the other for the analysis of their contents. The characteristics of the pirated content BTM should look for are defined by the user as a set of rules that are employed during the analysis of the collected data.

Studies that focus on a general characterization of the BitTorrent "universe" include the work of Zhang *et al.* [3], which analyzes torrents from five public communities through traces collected from trackers and DHT networks. Authors present, among other results: which are the main BT communities; the participation degree of each torrent publisher; the loads and localization of most used trackers; the geographic distribution of peers; and the most used BitTorrent implementations. Similarly, Zhang *et al.* [4] present an investigation about "darknets" in BT. These are private communities accessible only through subscription and the possible source of initial distribution of illegal copies. Among the results, authors compare characteristics of swarms promoted by darknets against ones from public communities.

Studies that focus on content dissemination in BT networks include the work of Blond *et al.* [5], which presents an analysis of 103 days of monitoring from three popular BT communities. Its results show a profile of the most active content providers and consumers. Authors were able to identify 70% of providers, list the most popular contents being shared and characterize the most active participants (users present in most swarms). Cuevas *et al.* [6] studies socio-economic factors from BT networks, emphasizing the incentives that drive content providers. Three groups of publishers are identified: those who distribute content due to financial incentives, those who act due to altruistic motivation and those who are responsible for fake content. Based on the analysis of one month of traces, the groups are characterized according to: the ISPs to which they are associated; types of content that are published; incentives for their activity and an estimation of possible monetary incomes.

The aforementioned studies from Zhang *et al.* try to quantify and model BT swarms through creation of "snapshots" of their lifecycle. Their scope, however, did not include an analysis about the patterns and dynamics of the dissemination of illegal copies of content. Considering the employed monitoring techniques and the results granularity, the studies by Blond *et al.* and Cuevas *et al.* are the ones most similar to the one presented in this paper. Results presented in these studies, however, do not present the necessary information to allow the identification of trends in the behavior of users that disseminate illegal copies of copyrighted content. Issues that directly influence the comprehension of processes used on the distribution of such copies are left unexplored. It is also important to note that these studies do not seem to consider technical issues such as the filtering out of polluted content, which is characterized by torrents with very short lifetime in the community. Such torrents should be carefully analyzed to guarantee that their content will not generate spurious results.

In this section we reviewed the most relevant studies that are related to the one presented in this paper. There are efforts from the P2P research community in order to create the necessary monitoring tools and proceed with observations to better understand the BT "universe". None of these studies, however, focused on understanding how BT networks are used for the dissemination of illegal copies of copyrighted content. Knowledge about these still unknown dissemination dynamics can support the development of effective strategies and mechanisms for the protection of copyrighted content. It can also contribute to stimulate the adoption of BT networks for commercial activities. We believe that such benefits are a sound reason to justify further investigation of the proposed topic. To the best of our knowledge, this is the first study that focuses on systematically mapping the dissemination process of illegal copies of content in BT networks. The following sections present the employed monitoring architecture, its instantiation and the most relevant results that were found.

## III. Monitoring Infrastructure

The volume of files shared in BitTorrent communities is huge [1]. To make our observations possible (about illegal distribution of copyrighted material), we implemented and instantiated a dedicated monitoring infrastructure. This enabled us to keep track of thousands of swarms. The resulting infrastructure allowed us to collect traces of 40,993 torrents, 915 trackers and more than 1.3 million IP addresses in a timespan of four months (from 05/2011 to 08/2011). The employed monitoring architecture, TorrentU [7], and the extensions developed in order to allow the required observations are presented in Sub-section III-A. Sub-section III-B presents information about the instantiation of the monitoring and the execution of our experiments.

### A. TorrentU Monitoring Architecture

TorrentU [7] is a flexible architecture designed and developed for monitoring BitTorrent networks. As presented in Figure 1, the architecture follows the classic manager/agent approach and thus basically contains two elements: an Observer and Telescopes. The observer acts as the manager of the architecture. It is a front-end that allows the operator to configure the system and observe the collected results in real time (and also historic data). Telescopes, in turn, act as agents. They are the components responsible for monitoring the BitTorrent universe and returning results according to requests received from the Observer. Telescopes are further divided in three components named "lenses", each one responsible for monitoring a different group of elements from the universe: communities, trackers and peers. Such modularization allows existing lenses to be changed and also new ones to be easily incorporated into the architecture without modification of other essential components.

Taking advantage of the flexibility provided by the TorrentU architecture, new functionality (not originally envisaged) was

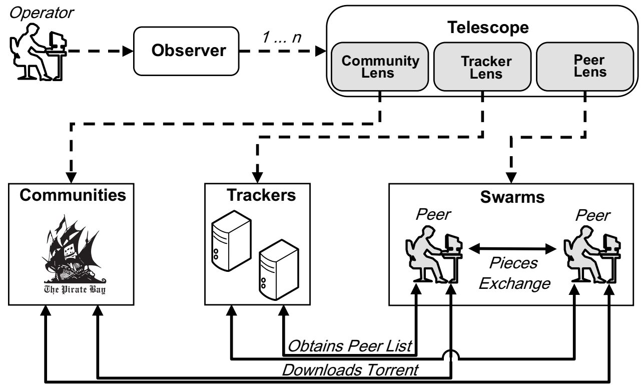



Figure 1. TorrentU architecture

implemented and integrated, resulting in two main extensions. The first, created to allow identification of the first seeders of a swarm, is a **seeker lens** that captures torrents as soon as they are published in a community. The second extension is a **torrent lens** that continuously monitors the community Web page (the ones containing information about the captured torrents) in order to collect: swarm lifetime; number of seeders and leechers; and comments posted by users about the content. The goal of the second extension is to create snapshots of the swarm throughout its lifetime.

Algorithm 1 presents a general view of the monitoring process. The first step consists in capturing recently published torrents from communities. For each torrent, a characterization of the tracker is performed and then it is contacted with a request for the peer list of the swarm. If the list is successfully received, a characterization of the first peers participating the swarm is performed. Next, the torrent lens is initialized for the processed torrent. It should be noted that no content is downloaded during the monitoring: peers are only contacted for acquisition of their bitfields.

There are six parameters that control the execution of the algorithm: *time* determines the duration of the whole monitoring campaign; *attempts* defines the number of connection attempts to trackers; *quantity* represents the size of the peer list requested to a tracker; *threshold* defines the number of peers of a swarm that will be contacted for acquisition of their bitfield; *frequency* defines the time interval between snapshots of a swarm; and *interval* represents the waiting time between each iteration of the algorithm.

It should be noted that this paper focus lies in the results of characterizing illegal movie copies dissemination, and not in the description of the monitoring architecture. The reader interested in more information about the architecture should refer to [7].

### B. Architecture Instantiation

Telescopes were deployed in three nodes of the PlanetLab testbed [16] and in a private server. The Observer component, in turn, was deployed on a single workstation.

Among existing open BitTorrent Communities, we chose PirateBay [17] due to its popularity. The Web pages of this community contain only links for torrents that are published through their servers. They also present statistics about users that published a torrent and provide user classification based on his/her reputation in the community.

```
input: time, attempts, quantity, threshold,
    frequency, interval
for i ← 0 to time do
    list[torrent] ← CaptureRecentTorrents();
    for j ← 0 to list[torrent].size() do
        torrent ← list[j];
        DownloadTorrent(torrent);
        ReadFile(torrent);
        CharacterizeTrackers(torrent);
        peerList ← GetPeerList(torrent,
        attempts, quantity);
        CharacterizePeers(peerList);
        if peerList.size() < threshold then
            ExchangeBitfields(torrent);
        end
        BeginSnapshotCapture(torrent,
        frequency);
    end
    Wait (interval);
end
```

**Algorithm 1:** Monitoring process

The parameters of the monitoring process (as previously described for Algorithm 1) were configured as follows: duration of monitoring (*time*): 4 months; attempts to contact a tracker (*attempts*): 2; number of peers requested from trackers (*quantity*): 50; number of peers contacted in a swarm for bitfield acquisition (*threshold*): 10; interval between snapshots of a swarm (*frequency*): 8 hours; waiting time between iterations of the algorithm (*interval*): 2 minutes.

## IV. Results

In order to avoid spurious data in our results, three filtering processes were employed to the 40,993 monitored torrents. First we removed all torrents for which all referenced trackers were inaccessible due to malformed URLs. This procedure filtered out 14,194 torrents. Next, we removed all torrents whose swarms could be contacted only in the first iteration of their monitoring. We observed that these swarms could not be further contacted because they were removed from the community. We assume that torrents which are almost immediately (*i.e.* under 8h) removed from the community by the administrators are invalid ones or contain polluted content. This filtering step eliminated 16,365 torrents. Finally, we removed all torrents whose trackers returned error messages upon contact. This final step filtered out 1,989 torrents. To the best of our knowledge, none of the previous studies related to ours [5], [6] employed a similar filtering process and possibly generated biased results with influences from spurious traces.

After the filtering process, there remained 8,445 torrents for investigation. Our study focused on four main analyses. First we present the characteristics of *producers*, who generate the illegal copies that will be distributed, and *publishers*, who make available the illegal copies in the PirateBay community. We focus on identifying their activity degree and possible relationships among them. Next, we present the most common digitalization processes applied to the observed files, demonstrating their influence in the publishing of copies through the lifetime of a movie. Third, we characterize the first seeders, who bootstrap the dissemination process, acting as initial content providers. Finally, we look into possible relationships among the activities of producers, publishers and providers. The characterization of dissemination from the consumers perspective, although deemed very important, has been left out of this investigation and will be addressed in a future work.

### A. *Producers and Publishers of Illegal Copies*

As mentioned in Section II-A, digitalization groups are responsible for the creation of illegal copies of the movies shared in BT networks. In this study, as previously noted, they are named producers. From the 8,445 analyzed torrents, 5,581 (66.08%) identified the producer that created the file. These copies were created by 482 distinct producers; the 10 most frequent ones are presented in Table II[1]. Figure 2 presents a Cumulative Density Function (CDF) in which the horizontal axis represents the producers ordered by volume of created copies and the vertical axis, the cumulative proportion of created torrents. This CDF shows that a small number of producers are responsible for most of the created files: almost 78% of the copies were created by 100 producers (20.74% of the 482 producers).

Table II
Content Producers Ranking

| Group | Torrents # | % |
|---|---|---|
| Dmt | 249 | 4.46 |
| Imagine | 189 | 3.38 |
| Mastitorrents | 158 | 2.83 |
| DutchTeamRls | 135 | 2.41 |
| Mtr | 134 | 2.40 |
| Extratorrentrg | 124 | 2.22 |
| Waf | 123 | 2.20 |
| Ddr | 109 | 1.95 |
| Miguel | 105 | 1.88 |
| NLT | 100 | 1.79 |

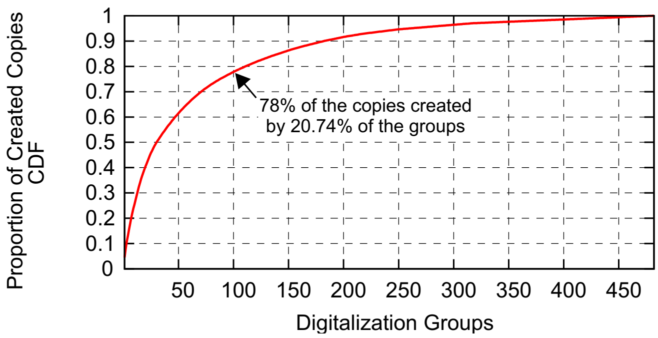



Figure 2. Groups cumulative contribution

After a torrent has been created by a producer, it may be published in a community. This step is executed by publishers.

[1]In our analysis, we present the pseudonyms of digitalization groups and community users to help illustrate our results and insights. It should be noted that we do not try to link these names to users real identities. Furthermore, the presented activity rankings are expected to change over time due to file sharing communities dynamicity. Such rankings, however, do not influence the presented results about the dissemination of illegal copies in file sharing communities.

These are, in the scope of this study and as mentioned earlier, registered users from the PirateBay community that uploaded the analyzed torrent files. This community divides its users in four categories (listed in descending order of privilege): VIP, trustworthy, helpful and normal (initial category of every user). The category of a user allows his/her reputation in the community to be inferred. Regarding the 8,445 analyzed torrents, they were published by 976 distinct users, assuming that each user holds a single identity. Table III presents the most active ones, their category in the community, and the number and proportion of published torrents. It should be noted that nearly all of the most active publishers listed in Table III are from categories with elevated privileges in the community. Figure 3 illustrates a CDF representing torrent publishers activity. The horizontal axis represents the cumulative number of publishers and the vertical axis the proportion of published torrents. This graph shows that few users are responsible for most of the published content: 125 users (12.8% of 976) published 76.17% of the content.

Table III
PUBLISHERS RANKING

| User | Category | Torrents # | Torrents % |
|---|---|---|---|
| .BONE. | VIP | 742 | 8.78 |
| MeMar | VIP | 261 | 3.09 |
| HDvideos | Regular | 236 | 2.79 |
| Black1000 | Trustworthy | 175 | 2.07 |
| SaM | VIP | 166 | 1.96 |
| virana | Trustworthy | 162 | 1.91 |
| sceneline | VIP | 146 | 1.72 |
| sadbawang | VIP | 128 | 1.51 |
| furtaperas | VIP | 111 | 1.31 |
| miguel1983 | VIP | 108 | 1.27 |

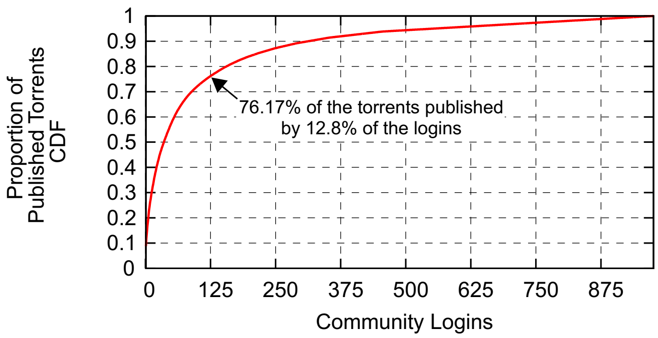



Figure 3. Publishers cumulative contribution

Figure 4 presents the relationship among producers (groups) and publishers (logins). The size of each circle illustrates the number of copies from a specific group that were published by one specific user of the community. Four typical cases were observed: (*i*) strong correlations between a producer and a publisher, as exemplified by case A, which illustrates the user "MeMar" publishing 236 of the 249 torrents produced by the group "Dmt"; (*ii*) a very active user publishing torrents from many groups, as exemplified by case B, in which user ".Bone." published torrents of 155 different groups; (*iii*) a producer supported by different publishers, as exemplified by case C, in which group "Axxo" has its copies published by users ".Bone." and "Test_Verify"; and (*iv*) digitalization group pseudonyms employed as community logins, whose are responsible for publishing a large number of torrents from the group of same name, as observed in D.

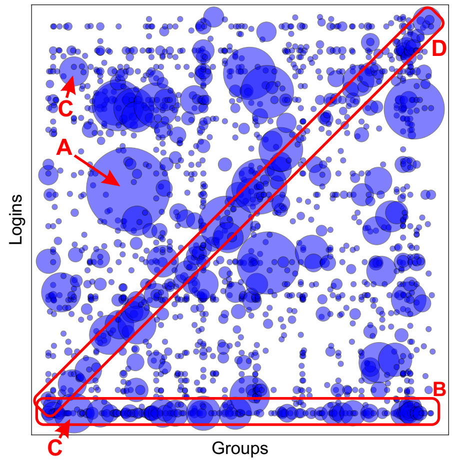



Figure 4. Relationships between producers (groups) and publishers (logins)

## B. Employed Digitalization Processes

Recall from Section II-A that distinct digital copies of a movie may have different qualities depending on the employed digitalization process. From the 8,445 analyzed torrents, 6,592 (78.05%) identified the process employed in the creation of the copy. Table IV presents a correlation between processes and groups, summarizing the processes, their degree of occurrence and the digitalization groups responsible for the greater number of copies within each process.

Table IV
MOST FREQUENT DIGITALIZATION PROCESSES

| Process | Torrents # | Torrents % | Main Groups |
|---|---|---|---|
| DVDRip | 5,138 | 77.94 | Dmt, Extratorrentrg, Mr_Keff |
| TS | 442 | 6.70 | Imagine, Dtrg, Feel-Free |
| DVDScr | 328 | 4.97 | Mastitorrents, Ddr, Teamtnt |
| R5 | 273 | 4.14 | Imagine, Cm8, Vision |
| CAM | 177 | 2.68 | Imagine, Feel-Free, Wbz |
| PPVRip | 94 | 1.42 | Iflix, Imagine, Dmt |
| TC | 75 | 1.13 | Team Tc, Mtr, Mastitorrents |
| SCR | 64 | 0.97 | Scr0n, Mastitorrents, Castellano |

Results show that the "DVDRip" process is by far the most common digitalization method, being employed in 77.94% of the copies. This prevalence may be explained by two factors. First, the quality of the copies generated using "DVDRip" process present the best quality among the considered types, so it is intuitive that it can attract more interest from users. Second, the DVDRip digitalization method is comparatively simpler and the media (DVD or Blu-ray discs) necessary for the process is widely accessible to the average user.

Other processes that stand out are "CAM", "TS", "DVDScr" and "R5". Their popularity reflects a trade-off among the difficulty of access to the source media for digitalization, the quality of this source and the time after the movie premiere in which it will be available. For example: we observed that the "TC" process presents the best quality among methods with releases expected within the first 4 weeks after the movie premiere. However, since processes "CAM" and "TS" require easily-accessible sources, they are more frequently employed (9.38% in comparison with 1.13% of the "TC" process). Another observed behavior is the specialization of certain groups in the creation of copies that employ processes based on sources that are hard to obtain. For example, the "Imagine" group is the second in number of created copies according to Table II. It is, however, the main producer of copies based on "TS", "R5" and "CAM" processes. This demonstrates the resourcefulness (and importance) of "Imagine" due to its early access to sources not easily obtainable.

The digitalization processes employed in the creation of copies can also be correlated to the lifecycle of movies after they premiered. Figure 5 illustrates the occurrence of copies of nine different movies, produced using various digitalization processes. Circle sizes represent the amount of torrents for each identified copy. The horizontal axis represents, in weeks, how long it took for the movie to be published after its premiere (according to IMDB [18]).

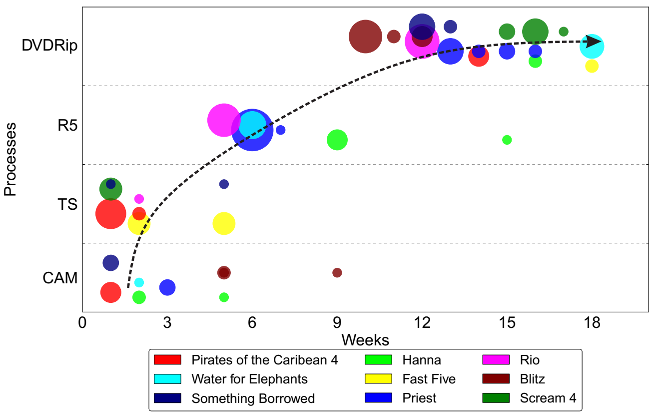



Figure 5. Digitalization processes according to weeks after premiere

Three aspects should be noted in the analysis of Figure 5. First, the publication of copies using a certain process tends to be concentrated in time, like a burst. This can be observed in the fifth week, when several copies of the movie "Rio" created using the "R5" process arise. Second, torrents containing a copy of a movie created using a specific digitalization process may be published even after the initial burst, as observed for the movie "Priest" over the 14th, 15th and 16th week. This behavior seems to occur due to specializations over previously published copies (such as employment of other codec types or the addition of new audio or subtitle languages). The third aspect is that some movies may not appear in all digital formats because of the absence of the corresponding source. Examples of this behavior are observed in "Fast Five" and "Pirates of the Caribbean 4", which did not have copies created with the "R5" process because, for example, the lack of a DVD region 5 source for these movies.

The data presented in Figure 5 was employed to generate the first approximation of a Markov chain representing the evolution of digitalization processes throughout a lifetime of a movie. The resulting model is illustrated in Figure 6. Three aspects can be highlighted: (*i*) in the beginning, "CAM" and "TS" processes are typically employed in the creation of the first copies, with the former being predominant; (*ii*) the recurrence encountered in the "CAM" state indicates that the process may be repeatedly employed, for example when a source of better quality is obtained by producers; and (*iii*) the digitalization process of a given movie will eventually converge to "DVDRip", in which case other processes are not used in new copies.

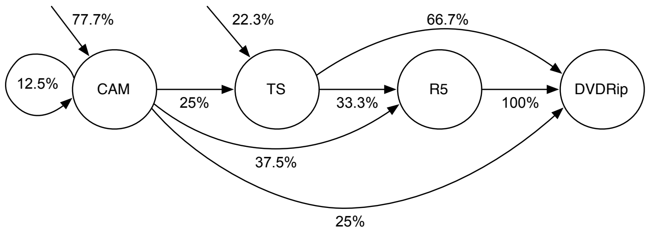



Figure 6. Markov chain of digitalization processes evolution

### C. *Providers of Illegal Copies*

The dissemination of illegal copies in BT networks depends on providers. These are, as previously mentioned, users with the first copy of the file. Without these providers, dissemination would not be possible. Recall that, in BitTorrent, the user who publishes the torrent in the community is not necessarily the one in possession of the file containing the copied content. These users, also known as first seeders, are responsible for bootstrapping the swarm. To characterize these users, we analyzed the peerlist obtained from trackers early in swarms lifetime.

Torrents added to the community were captured as soon as they were detected by the TorrentU community lens (recall from Section III-B that the community is probed in intervals of 2 minutes). Once detected, the corresponding torrent trackers were contacted for receipt of the initial peerlist. The shorter the time between a torrent publication and receipt of its peerlist, the higher the chances that only the first seeders will be contained in the tracker response. In our study this interval was 4 minutes on average.

We begin our analysis by characterizing each contacted tracker. From 915 observed trackers, 150 (16.39%) answered the query with a valid peerlist. We were able to identify the geographical location of 141 trackers. Figure 7 illustrates the obtained results. In the graph, each bar represents the number of trackers found in a specific country and its color the respective continent. Results indicate that Europe stands out as the location of most trackers with a total of 64 hosts, 37 of which are on the Netherlands. North America appears in the second position, due to the United States, which hosts 36 trackers. Finally, Asia appears as the third continent in number of trackers. These specific results are overall in line with other general ones previously obtained by Zhang *et al.* [3].

After contacting the trackers, the obtained responses were analyzed and led to the identification of the first seeders in 4,235 (50.14%) of the torrents. These were seeded by 5,227 peers (a few swarms presented more than one initial seeder). These peers were associated to 1,887 unique IP addresses. Table V presents the list of most active users, indicating their country of origin, ISP, and number of swarms joined as initial seeder. Figure 8, in turn, show a CDF of the cumulative

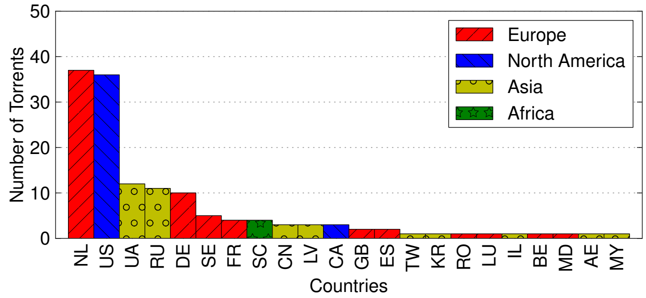



Figure 7. Geographical location of trackers

contribution of first seeders according to the proportion of analyzed torrents.

From Table V and Figure 8, two insights can be obtained from the analyzed results. The first is that 5.29% of the observed IP addresses (100 of 1,887) participated as first seeders of 56.51% of the analyzed swarms. This result indicates the possibility that specialized users are employing *seedboxes* [19] in order to disseminate their content. The second aspect is that 79.33% of the IP addresses (1,497 of the 1,887) exclusively seeded one or two swarms. Such behavior indicates that these users may be "domestic" ones, sharing illegal copies of very specific content types.

Table V
SEEDER RANKING

| Country | ISP | # Swarms |
|---|---|---|
| FR | Ovh | 408 |
| NZ | Obtrix | 186 |
| US | - | 119 |
| FR | Ovh | 104 |
| ES | - | 90 |
| FR | Ovh | 86 |
| PL | Mokadi | 85 |
| GB | Uk-Ovh | 79 |
| FI | Lsinki | 75 |
| NL | Ziggo | 70 |

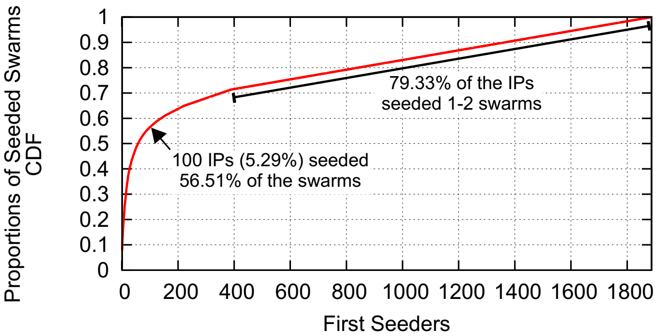



Figure 8. Providers cumulative contribution

In order to identify the location of the observed content providers, we considered the first seeder of each swarm a unique entity, even if peers from distinct swarms presented the same IP address. We successfully determined the location of 5,128 of the 5,227 observed seeders. Figure 9 shows the 26 countries presenting higher concentration of first seeders (countries not presented in the graph always contained less than 20 seeders). Results indicate that Europe stands out as the most common location of content providers, hosting nearly four times more first seeders (63.39%) than Asia (17.25%), which appears in second place. Another behavior observed is that France, United States and the Netherlands present considerable higher number of first seeders than the average measured for the other countries.

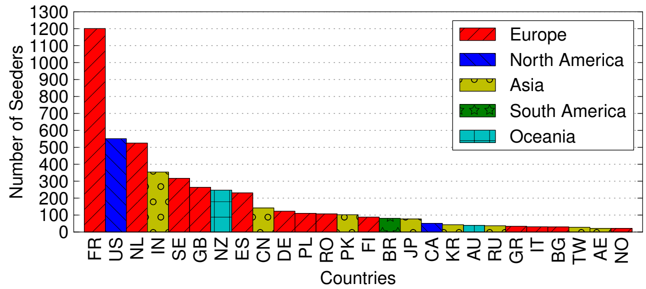


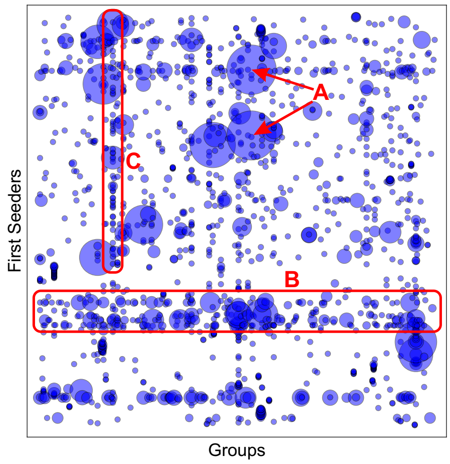

Figure 9. First seeders location

Through the characterization of providers (first seeders), producers (digitalization groups) and publishers (community users) we identified examples that indicate the existence of relationships among these entities. Figure 10(a) presents the correlation of digitalization groups and first seeders. Three points of interest are highlighted: First, some providers are dedicated to the dissemination of specific producers copies. This can be observed in A, which represents the group "Miguel", which had 93.68% of its swarms seeded by two IP addresses only. Second, some providers serve copies of various producers, as observed in case B. Third, a producer may be served by a diverse group of providers. This may be observed in case C, which represents the group "Dmt", which had its copies provided by 62 different seeders.

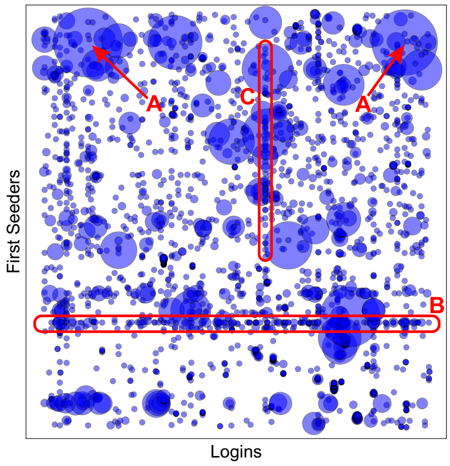
Figure 10(b) characterizes associations between community logins and first seeders. Three points of interest are highlighted. First, providers with high degree of activity may be associated with a single community login. This can be exemplified by case A, in which two specific IPs seed copies published by users "black1000" and "virana". Second, one provider may serve a diverse group of publishers, as in case B. Finally, one publisher may be served by a diverse group of providers. This corresponds to case C in the figure, representing a user ("MeMar") who had its published copies seeded by 61 unique IPs.

## V. Conclusions and Future Work

Content sharing through BitTorrent networks is one of the activities that generate most of the Internet traffic. It is known that most of this traffic is related to the sharing of illegal copies of various types of copyrighted content. So far, even with the relevance of this research topic, no studies related to the characterization of BitTorrent content sharing have focused on mapping the dissemination dynamics of illegal copies. Aiming at bridging this gap, we presented a detailed study of traces registering four months of activities from one of the most, if not the most, popular BitTorrent file sharing community. Traces were collected with an extension of a BitTorrent




(a) Groups × first seeders




(b) Logins × first seeders

Figure 10. Relationships among producers (groups), publishers (logins) and providers (first seeders)

monitoring architecture designed to observe the BT "universe". The analysis of the collected data allowed us to obtain new insights about the dissemination of illegal copies of content. To the best of our knowledge, this is the first scientific study that focuses on characterizing the dissemination of illegal copies of content in BitTorrent networks.

Based on the obtained results it is possible to identify behavior patterns of the sources distributing illegal copies of content. Regarding their producers, it was found that most torrents present an identification of the digitalization group responsible for its creation and that most copies are generated by a small number of groups. In the case of publishers, a behavior similar to the one exhibited by producers was observed, in the sense that most torrents are published by a small number of active users. An association between producers and publishers was also identified. Analyzing the employed digitalization processes, we discovered that certain producers are specialized in specific processes. Our study helps understanding and quantifying the evolution of the digitalization processes employed throughout the lifetime of a movie after its premiere. Finally, by analyzing the peers responsible for the initial seeding of illegal copies, we identified relationships among producers, publishers and providers. The results are relevant for operators of Internet and media service providers, the film industry and, very important, to researchers of this community involved in designing effective models and mechanisms to securely operate large-scale content delivery solutions.

During our study, we identified three opportunities of future work. The first one consists in observing the behavior of users from a larger number of BT communities. A second opportunity is the observation of "darknets": private BT communities that might be the starting point for the dissemination of illegal copies of contents later found on public communities. Finally, it would be interesting to observe content consumption dynamics and patterns. Examples are: characterization of leechers geographical location, which may indicate regional trends in content consumption; and the characterization of migration patterns of leechers among swarms.

## REFERENCES


[1] H. Schulze and K. Mochalski, "Internet study 2008-2009," 2009, available at http://www.ipoque.com/sites/default/files/mediafiles/documents/internet-study-2008-2009.pdf, access in July 2011.

[2] Envisional, "An estimate of infringing use of the internet," 2011, available at http://documents.envisional.com/docs/Envisional-Internet_Usage-Jan2011.pdf, access in July 2011.

[3] C. Zhang, P. Dhungel, D. Wu, and K. Ross, "Unraveling the bittorrent ecosystem," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 7, pp. 1164–1177, 2010.

[4] C. Zhang, P. Dhungel, D. Wu, Z. Liu, and K. Ross, "Bittorrent darknets," in *INFOCOM 2010, IEEE International Conference on Computer Communications*, 2010, pp. 1460–1468.

[5] S. Le Blond, A. Legout, F. Lefessant, W. Dabbous, and M. A. Kaafar, "Spying the world from your laptop: identifying and profiling content providers and big downloaders in bittorrent," in *LEET 2010, USENIX Workshop on Large-Scale Exploits and Emergent Threats*, 2010, pp. 4–4.

[6] R. Cuevas, M. Kryczka, A. Cuevas, S. Kaune, C. Guerrero, and R. Rejaie, "Is content publishing in bittorrent altruistic or profit-driven?" in *Co-NEXT 2010, International Conference on Emerging Networking Experiments and Technologies*, 2010, pp. 1–12.

[7] R. B. Mansilha, L. R. Bays, M. B. Lehmann, A. Mezzomo, L. P. Gaspary, and M. P. Barcellos, "Observing the bittorrent universe through telescopes," in *IM 2011, IFIP/IEEE International Symposium on Integrated Network Management*, 2011, pp. 1–8.

[8] F. R. Santos, W. C. Cordeiro, L. P. Gaspary, and M. P. Bracellos, "Choking polluters in bittorrent file sharing communities," in *NOMS 2010, Network Operations and Management Symposium*, 2010, pp. 559–566.

[9] Y.-J. Lee, J.-H. Jeong, H.-Y. Kim, and C.-H. Lee, "Energy-saving settop box enhancement in bittorrent networks," in *NOMS 2010, Network Operations and Management Symposium*, 2010, pp. 809–812.

[10] X. Fan and Y. Xiang, "Modeling the propagation of peer-to-peer worms under quarantine," in *NOMS 2010, Network Operations and Management Symposium*, 2010, pp. 1542–1201.

[11] H. Lee, A. Nakao, and J. Kim, "Traffic control through bilateral cooperation between network operators and peers in p2p networks," in *NOMS 2010, Network Operations and Management Symposium*, 2010, pp. 583–590.

[12] Wikipedia, "List of warez groups," 2011, available at http://en.wikipedia.org/wiki/List_of_warez_groups, access in July 2011.

[13] K. Bauer, D. McCoy, D. Grunwald, and D. Sicker, "Bitstalker: Accurately and efficiently monitoring bittorrent traffic," in *WIFS 2009, IEEE Workshop on Information Forensics and Security*, 2009, pp. 181–185.

[14] K. Junemann, P. Andelfinger, J. Dinger, and H. Hartenstein, "Bitmon: A tool for automated monitoring of the bittorrent dht," in *P2P 2010, IEEE International Conference on Peer-to-Peer Computing*, 2010, pp. 1–2.

[15] K. Chow, K. Cheng, L. Man, P. Lai, L. Hui, C. Chong, K. Pun, W. Tsang, H. Chan, and S. Yiu, "Btm - an automated rule-based bt monitoring system for piracy detection," in *ICIMP 2007, International Conference on Internet Monitoring and Protection*, 2007, pp. 1–6.

[16] PlanetLab, 2011, available at http://www.planet-lab.org, access in July 2011.

[17] Piratebay, 2011, available at http://thepiratebay.org, access in July 2011.

[18] IMDB, 2011, available at http://www.imdb.com, access in July 2011.

[19] Wikipedia, "Seedbox," 2011, available at http://en.wikipedia.org/wiki/Seedbox, access in July 2011.