Movie Pirates of the Caribbean: Exploring Illegal Streaming Cyberlockers

Damilola Ibosiola,[†] Benjamin Steer,[†] Alvaro Garcia-Recuero,[†]
Gianluca Stringhini,[‡] Steve Uhlig,[†] Gareth Tyson[†]
[†]Queen Mary University of London, [‡] University College London
[†]{d.i.ibosiola, b.a.steer, alvaro.garcia-recuero, steve.uhlig, g.tyson}@qmul.ac.uk, [‡]g.stringhini@ucl.ac.uk

Abstract

Online video piracy (OVP) is a contentious topic, with strong proponents on both sides of the argument. Recently, a number of illegal websites, called *streaming cyberlockers*, have begun to dominate OVP. These websites specialise in distributing pirated content, underpinned by third party indexing services offering easy-to-access directories of content. This paper performs the first exploration of this new ecosystem. It characterises the content, as well the streaming cyberlockers' individual attributes. We find a remarkably centralised system with just a few networks, countries and cyberlockers underpinning most provisioning. We also investigate the actions of copyright enforcers. We find they tend to target small subsets of the ecosystem, although they appear quite successful. 84% of copyright notices see content removed.

1 Introduction

Online Video Piracy (OVP) has been the focus of an increasing debate over the past years. Entire political movements have emerged around the idea that content should be freely available (Miaoran 2009), whilst lobbyists consistently argue that dire consequences exist. For example, CBP reported that piracy costs the US economy over 750,000 jobs, and between $200-250B per year (Raustiala and Sprigman 2012). Regardless of one's stance, it is undeniable that OVP constitutes a major web traffic generator (Monitor 2011; Elder 2016), and creates significant interest from users, law enforcers and the creative industries alike.

Traditionally, online piracy was dominated by decentralised peer-to-peer (P2P) systems such as Gnutella and BitTorrent. However, these have since been surpassed by a new breed of more centralised service allowing users to stream pirated content directly from YouTube-like websites — so called *streaming cyberlockers*. These streaming cyberlockers have gained huge traction. For example, many prominent portals are in the Alexa Top 1K, *e.g. openload.co*, *thevideo.me* and *vidzi.tv*. Their ease of use attracts a large number of users and the difficulties law enforcers encounter when detecting user identities provides viewers with relative safety from prosecution. Organisations such as the Motion Picture Association of America (MPAA) have, therefore, shifted their efforts towards shutting down the cyberlockers themselves. Examples of prominent shutdowns witnessed in this paper include *allmyvideos.net*, *vidbull.com* and *vodlocker.com*.

Although similar to typical social video platforms, these streaming cyberlockers address a very different need. They employ few, if any, copyright checks and utilise evasion tactics to avoid detection. For example, they often curate content on their front-pages to appear legitimate and disable search to prevent visitors from looking up videos. This has created an interesting ecosystem where cyberlockers depend on third party (crowd-sourced) *indexing websites* that create a searchable directory of direct links (URLs) to the videos. These two types of website operate hand-in-hand with a symbiotic relationship, collectively underpinning a global network of online piracy.

To date, little is known about this emerging ecosystem. Its exploration, however, could reveal a range of insights regarding how large-scale copyright infringement takes place. This raises several particularly interesting questions, including: what type of copyright content is shared? What are the dynamics regarding both content and website appearance/disappearance? What web hosting characteristics are commonly seen and how resilient are they? How are these websites pursued by copyright enforcers and how do the websites react?

To answer these questions we exploit several measurement methodologies (§3), acquiring evidence of the characteristics exhibited in this domain. As it would be impossible to inspect the entire copyright infringement ecosystem, we have taken a slice of 3 prominent indexing sites, as well as 33 different cyberlockers. Between January and September 2017 we performed monthly crawls, collecting all published videos on these indexing sites. In parallel, we have scraped their related cyberlockers, collecting data on each video, including its availability and where it is hosted. To complement this data we further gathered metadata on the videos themselves, *e.g.* release date and genre. Finally, we have monitored legal take down notices, allowing us to understand the reaction of the cyberlockers to complaints.

We begin our analysis by exploring the streaming links shared on indexing sites (§4). We find a set of web platforms actively involved in aggressive copyright infringement. Predominantly content is made up of recently released Drama, Comedy, Thriller, and Action films. However, we also ob-

serve a non-negligible amount of older content — some videos are from over 100 years ago. The websites we monitor show clear temporal trends with periods of activity, followed by collapse — likely driven by legal take downs. For example, *putlocker.is* (an indexing site) ceased uploading new links three months into our measurements. This reveals a model rather more vulnerable than the decentralised P2P networks.

We then inspect the characteristics of the individual cyberlockers (§5). We model these concepts as several graphs that capture the related attributes of websites. A key finding is the apparent centralisation of these portals, with a small set of dependencies vulnerable to attack from copyright enforcers. For example, we observe that 58% of all videos are located within just two hosting providers (despite being spread across 15 cyberlockers). Similarly, we find strong signs that individual pirates tend to operate *multiple* websites. For instance, although seemingly different cyberlockers, *daclips*, *gorillavid* and *movpod* are all operated by the same owner. These three cyberlockers alone host 15% of observed content. Again, this suggests a distribution model that is far less resilient than its decentralised P2P counterparts.

Finally, we inspect the behaviour of copyright enforcers (§6). By studying the takedown notices placed against the cyberlockers under observation, we find that most enforcers take a bulk approach — selecting a set of cyberlockers and generating many notices. That said, most cyberlockers *do* appear to placate such enforcers. During our measurement period, 84% of notices later saw the content removed. Our results have implications for understanding modern copyright infringement both from the perspectives of content pirates and law enforcers (§7).

2 Background & Related Work

Before beginning our analysis, we provide a brief overview of the the general area, as well as related works.

2.1 Overview of Video Piracy Stakeholders

There are three major stakeholders worth considering. The failure of any of them would result in the collapse of the ecosystem. The players are:

Video Uploader: A video uploader harvests video content (*e.g.* using BitTorrent) and uploads it to a streaming cyberlocker. For each video uploaded, a unique URL is received. These URLs are published by the uploader on an indexing site with the appropriate metadata for searching.

Streaming Cyberlocker: A streaming cyberlocker is a web platform where a video uploader stores content. Typically a streaming site is neither searchable nor indexed by search engines. Users require the specific URL to view the content.

Indexing Site: Indexing sites operate as a public directory, mapping video metadata (*e.g.* title) to a list of cyberlocker URLs where the content can be viewed. They allow viewers to search for any desired video and select a preferred streaming site.

2.2 Related Work

Online video distribution is not a new topic. The streaming cyberlockers work on a model of third parties uploading content. There are a range of video platforms allowing users to upload and share their own content, *e.g.* YouTube (Zink et al. 2008; Torres et al. 2011; Cha et al. 2007) and Vimeo (Sastry 2012). Ding *et al.* characterised YouTube uploader behaviour and classified the uploads (Ding et al. 2011). It was discovered that the majority of content was copied and little actually user generated. Of most relevance to our work is the use of such platforms to distribute copyrighted material. There have been several studies looking at how platforms have been exploited for such purposes (Clay 2011; Hilderbrand 2007). In response, platforms like YouTube now employ signature-based detection to prevent copyrighted material remaining online (Dutta and Patel 2008). This has led to a range of unusual evasion techniques, *e.g.* removing portions of the film and injecting artefacts.

This complexity has resulted in pirated content moving away from these portals towards what are known as cyberlockers or one click file hosts (OCFH). These services offer remote storage, allowing users to share files. (Mahanti et al. 2012) provide an understanding of the nature of OCFHs and their effect on the network. Sanjuàs-Cuxar *et al.* also analysed HTTP traffic emanating from OCFHs, ranking them amongst the major contributors of HTTP traffic on the Internet (Sanjuàs-Cuxart, Barlet-Ros, and Solé-Pareta 2012). Perhaps closest to our own work is (Lauinger et al. 2013b; 2013a; Farahbakhsh et al. 2013). The first works scraped data from several OCFHs, such as MegaUpload and RapidShare, to understand the fraction of files that infringe copyright, whilst the second work investigated the impact of the MegaUpload shutdown on BitTorrent. Although closely related, our focus is not on *file sharing* but on pirated *video streaming*. We know of only one work targeting streaming services (Rafique et al. 2016). This work investigated the security implications of illegal sports streaming, as well as how deceptive adverts and malware are used for monetisation. These sports sites are quite different to the movie sites we observe, primarily because they are *live* broadcasts. Hence, we proceed to study the broader aspects of video piracy. Our paper sheds light on the behaviour of these websites in reaction to legal action, as well as the individual characteristics and relationships between them. To the best of our knowledge, this is the first paper focusing on the streaming cyberlocker ecosystem.

3 Methodology & Data Collection

We begin by presenting our measurement methodology. Our measurements follow three steps: (*i*) Collecting all streaming links from the indexing sites; (*ii*) Visiting the links to check the availability of the videos; and (*iii*) Gathering extended metadata for each video and website under study.

3.1 Indexing Sites

Due to the sheer number of indexing websites, it is impossible to evaluate them all. Hence the first step is to select a subset of indexing sites — these operate as "seeds" which

Indexing Site	No. of indexed videos	No. of videos with streaming links	No. of streaming links	% of videos with streaming links	No. of unique cyberlockers
putlocker.is	25,700	24,974	148,878	97.2	104
watchseries.gs	49,614	49,522	300,296	99.8	125
vodly.cr	64,021	55,313	346,524	86.4	84
Total	139,335	129,809	795,698	93.2	151

Table 1: Summary of data collected from each indexing site.

allow us to identify key cyberlockers. To achieve this, we inspected court orders obtained by the MPAA to understand those sites viewed as important by copyright enforcers. We then complemented this by performing a variety of searches on Google using relevant terms (*e.g.* "free films", "watch movies free"). This was intended to discover websites that a typical user may encounter when searching for free content. This is confirmed by industrial reports that highlight many of the cyberlockers we observe as key offenders (NetNames 2014). From these two data sources, we identified three regularly occurring websites: putlocker.is, watchseries.gs and vodly.cr (Orlowski 2013). These three sites mainly index streaming links to movies, with an additional small fraction of TV shows. In this paper, we use the term $video$ to refer to both. We emphasise that these may not be representative of *all* indexing sites — our analysis is specific to these three large sites, although we note these are significant players in the broader ecosystem.

We have designed a crawler that iterates over all video pages indexed on each of the three indexing sites. It extracts the video title, release year, genre and all associated streaming links. As previously stated, the indexing sites do *not* host any content — only links to external cyberlockers. We initiated this crawl on 12/01/2017 and repeated it on a monthly basis until 12/09/2017[1]. Table 1 summarises the data for each indexing site targeted.

3.2 Streaming Cyberlockers

After each monthly snapshot was gathered from the three indexing sites, the crawler followed each streaming URL to gain data from the cyberlockers themselves. We identified a total of 151 streaming cyberlockers on the indexing websites. We identify individual cyberlockers using their domain name; note that this includes mirrored cyberlockers with different Top Level Domains (TLDs). Unless stated otherwise, we treat these as different portals. The cyberlockers had diverse setups, and many had taken steps that made crawling challenging. For example, six domains used Dean Edward's compression algorithm[2] for obfuscating the server hosting the content. As it was impossible to scrape all 151 cyberlockers,[3] we selected the 33 most prominent streaming domains; this set covered 59.3% of extracted streaming links. The selected domains were those which were currently online and made the video information available to collect. The domains that were not selected were either offline at the time of scraping or redirected to a different site. Unlike YouTube, we found that the user interfaces were quite primitive, lacking reliable metadata *e.g.* view count and date of upload. For instance, 64% of examined streaming cyberlockers did not allow searching and 42% of portals "curated" their frontpages with legal short videos, which appear to have fake view counters. Therefore, for each video, we only recorded whether or not the video was online and the domain of the server it was hosted on.

3.3 Cyberlocker Metadata

Once we had collected all cyberlockers, we compiled metadata for each one. For every cyberlocker domain we performed DNS propagation checks around the world to generate domain $\rightarrow$ IP address mappings. We discovered a total of 1,903 distinct IP addresses hosting videos. We mapped each IP address into its geographical locations using Maxmind GeoLiteCity[4] and Autonomous System (AS) using Team Cymru.[5] We discovered servers distributed across 8 countries, 2 continents and 9 distinct Autonomous Systems (ASes). Following this, we loaded all cyberlocker homepages using phantomjs.[6] Upon each load, we recorded all the first and third party domains loaded by the page.

3.4 Lumen Database

A major theme in our work is understanding the role that video portals play in copyright infringement. It is, therefore, necessary to obtain ground truth data on which videos compromise copyright. To gather such data, we have scraped the Lumen database between 01/01/2017 and 30/09/2017 (the same period as our cyberlocker crawls). Lumen is a platform that aggregates legal complaints and requests for removal of online content. Each record covers an individual complaint to one or more organisations. An entry contains the URL(s), the complainant, the date and the complaint target (*i.e.,* a cyberlocker). Lumen predominantly captures complaints made to Google for removing content links from search results. Beyond this, Lumen also contains complaints to other search and social media sites, *e.g.* Bing and Twitter.

[1] *putlocker.is* was crawled for monthly period starting 12/01/2017, 12/02/2017, 12/03/2017 as it went offline afterwards. In the case of *vodly.cr*, we crawl it from 12/04/2017 onwards

[2] http://dean.edwards.name/weblog/2007/08/js-compression/

[3] Partly due to the frequency by which these websites change their web interface

[4] https://www.maxmind.com/

[5] http://whois.cymru.com

[6] http://phantomjs.org/

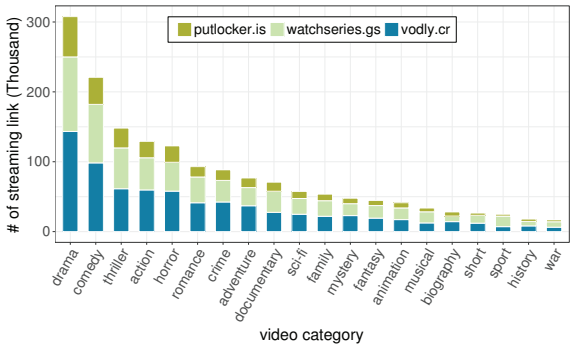

Figure 1: Number of streaming links per category.

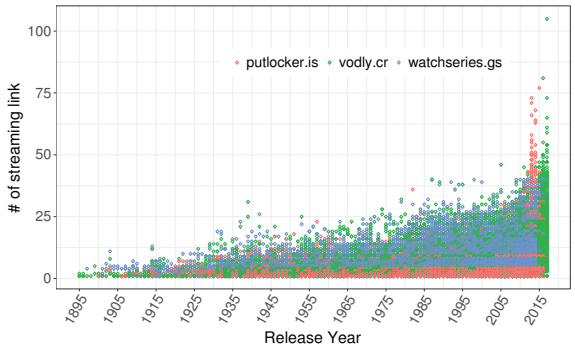

Figure 2: Number of streaming links per video page. Video pages are split into release year.

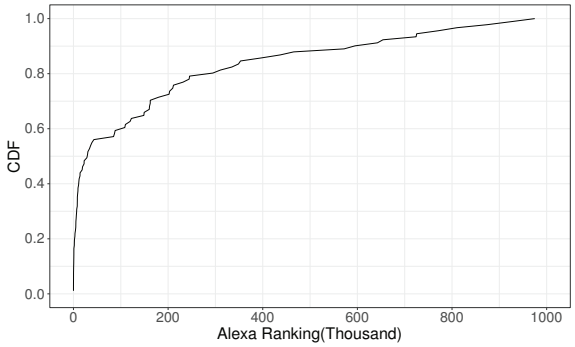

Figure 3: Cumulative Distribution Function (CDF) showing the distribution of streaming domains based on their Alexa Rank.

4 Characterising Indexing Portals

When a user wishes to view a video, the first entity they must interact with is an *indexing* site. In this section, we review what links are made available on these indexing portals, as well as the cyberlockers and content they point to.

4.1 How Many Links Are Available?

We begin by inspecting the *number* of content items being indexed over time. This can be measured from two perspectives: (*i*) the number of video pages made available (there is one page per video) and (*ii*) the number of streaming links made available on those pages. The former represents the number of new videos added to the indexing portals, whilst the latter captures the number of links per video. To give a brief understanding of the *types* of videos available, Figure 1 shows the number of links within the top 20 genres specified on the indexing sites (this also coincides with IMDB's[7] top 20 genres). It can be seen that Drama, Comedy, Thriller, Action and Horror videos dominate; the distribution in each indexing site is roughly equal and all follow an identical ranking.

When combining all genres we discover a total of 139,335 video pages and 795,698 streaming links. Figure 2 plots the number of streaming links attached to each video for each indexing site (across each release year). On average a video has 6 streaming links, but there is clear relationship between the recency of the release and the number of streaming links available. About 73% of links are for videos released since 2000. Diversity can also be observed across the different portals: this figure is 81% for *putlocker.is*, 74% for *vodly.cr* and 69% for *watchseries.gs*. This indicates that the portals offer different styles of corpora. Overall, the average number of streaming links for videos with recent release years ($\geq$2000) is 7, compared to just 4 for earlier releases.

We also observe that 7% of video pages list *no* streaming links; this suggests that either the links were removed, or the pages were generated without links being added. This is particularly prevalent for older videos. About 11% of videos with release years before 1980 do not have any streaming links, compared to just 6% for later release years. Only 0.3% of videos in 2017 have no links. This is likely driven by the higher demand and the more proactive participation of people uploading fresh content. That said, these portals also contain extremely old content, some over 100 years old. Characterising these portals as exclusive copyright-infringement platforms may therefore be misplaced. Curiously, the fraction of films released before 1950 without streaming links is actually lower than later films — just 6%. We assume this is because such videos are not aggressively pursued by copyright enforcers, hence reducing the number of legal actions.

[7]http://www.imdb.com/

4.2 Which Cyberlockers Are Most Popular?

The previous section inspected the *number* of streaming links. Next, we investigate *which* cyberlockers are most prominent. From the 795,698 streaming links extracted, there are 151 unique streaming cyberlocker domains. We first inspect their popularity as measured by the Alexa Rankings. Figure 3 presents a Cumulative Distribution Function (CDF) of the Alexa ranks for the cyberlockers. About 60% of these streaming domains are in Alexa's Top 1M. Amongst these, 70% are in the Top 200K. The top three most popular streaming sites are *openload.co* (rank 147), *thevideo.me* (543) and *vidzi.tv* (745). These rankings, however, do not correlate well with the number of videos hosted on the domain (Spearman coefficient of -0.015). For exam-

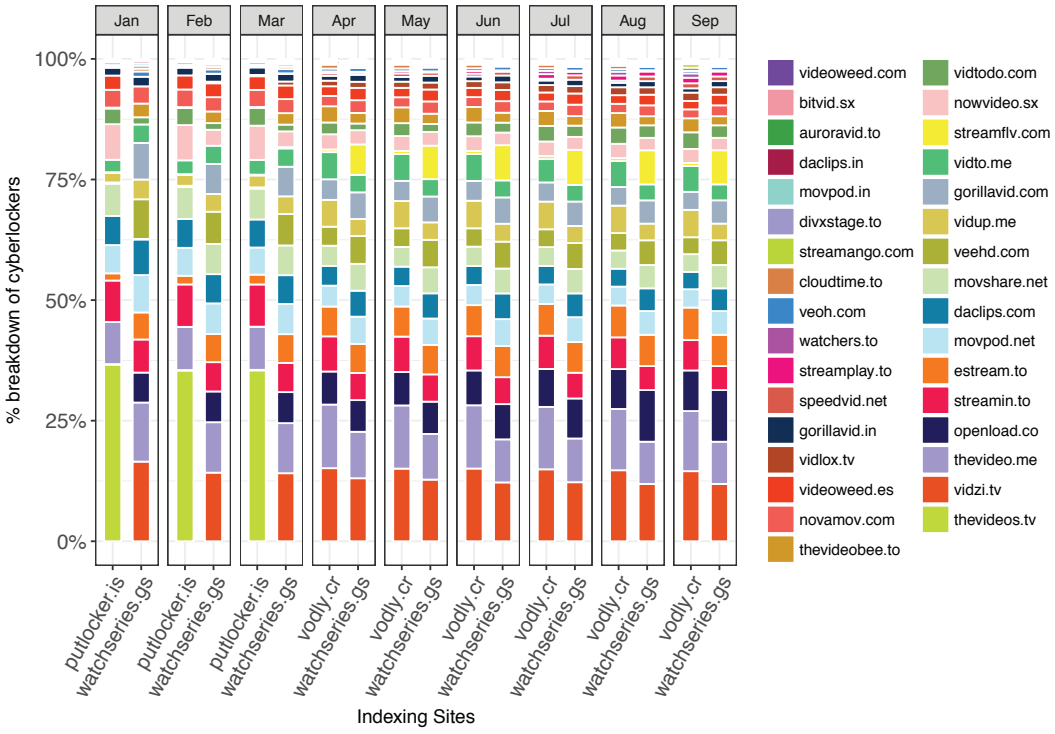

Figure 4: Breakdown of streaming links seen on each indexing site per month. We began crawling indexing site *vodly.cr* in April when *putlocker.is* was taken offline. The stacked bar is ordered with the largest cyberlocker at the bottom.

ple, *streamin.to* hosts 30,401 videos compared to just 7,288 for vidlox.tv and 1,924 for *streamango.com*. Despite this, the latter two rank 5,699 and 2,124 compared to just 6,625 for *streamin.to*. We can also inspect popularity through the lens of the indexing sites. Figure 4 presents a breakdown of the streaming links that make up the indexing sites, split by monthly snapshot. This is primarily intended to visualise the breakdown of cyberlockers per month, rather than their evolution over time. Note that the indexing sites vary across the time periods because *putlocker.is* ceased uploading new content in April, to be replaced by *vodly.cr*.

Firstly, it can be seen that well known user-generated content platforms such as YouTube, Vimeo or Dailymotion are not observed once. Instead, the indexing portals exclusively link to videos hosted on platforms that operate outside of the "mainstream", *e.g. thevideos.tv*, *movpod.net* and *videoweed.es*. Secondly, it can be seen that the cyberlockers present on each indexing site are different. This suggests communities where individual cyberlockers are associated with particular indexing sites. 30% of cyberlockers are exclusive to a single index; 33% are seen on two; the remainder appear on all indexing sites. The latter are, unsurprisingly, those with the greatest number of links. From the cyberlockers found on multiple indexing sites, 73% of their links are unique and seen once. In other words, only 27% of cyberlocker links are posted on more than one of our indexing sites. This suggests that different pirates have quite different strategies for promoting links to their content.

The prominence of each of these cyberlockers also changes across the monthly snapshots. For example, in February, we witness the introduction of *vidlox.tv* and *streamplay.to*; in March — *streamflv.com*; in April — *speedvid.net*; in July — *watchers.to* and in August — *streamango.com*. We also observe removals of cyberlockers, *e.g.* in April, *thevideos.tv* ceases to be indexed. This is because, prior to this, it was exclusively indexed by *putlocker.is*. Upon ceasing operation in April, the loss of *putlocker.is* meant that *thevideos.tv* disappeared from our vantage point.

We also see arrival and removal dynamics within individual links to each of the cyberlockers. Out of the 33 streaming cyberlockers we examined, we observed that 25 had links *both* added and removed. The remaining 8 had only additional links injected, and never had any removed: these were *openload.co*, *vidtodo.com*, *vidup.me*, *estream.to*, *streamplay.to*, *vidlox.tv*, *streamango.com*, *watchers.to*. In total 55% of cyberlockers saw growth during our measurement period, whilst 45% saw a decline. The most extreme was *divxstage.to* ,which in June had 24% of its links removed from the indexing sites. In contrast, in July *streamplay.to* saw a 107% increase in the number of links indexed. These aggressive dynamics are presumably enabled by the ease that uploaders can move between cyberlockers.

5 Characterising Cyberlockers

The above has revealed a wide range of cyberlockers. We next explore the characteristics of these cyberlockers, specif-

ically regarding (*i*) the use of third party domains within the webpages; (*ii*) the hosting of their video content; and (*iii*) the similarities between the webpage HTML structure. Whereas the first two aspects shed light on the design and build of the websites, the third provides insight into potential relationships that may exist between the websites.

5.1 Modelling Cyberlockers

To model the relationships between cyberlockers, we embed them into a series of graph structures. Each graph captures shared characteristics and potential relationships between the websites. We use several techniques to generate three graphs from our data:

- $Domains = (V, D, E_1)$, where V is the set of cyberlockers, D are third party domains, and E_1 link third party domains to the cyberlockers where they are embedded (in the homepage). As we are interested in identifying relationships, we filter domains with a degree of one.
- $Networks = (V, N, E_2)$, where N consists of Autonomous Systems (ASes), and E_2 is a set of directed links indicating that a cyberlocker is hosted within a given AS. This allows us to reason over the hosting strategies of these operators.
- $HTML = (V, E_3)$, where E_3 contains weighted links based on the homepage HTML similarity between two cyberlockers $\in V$. Similarity is computed using the tag-based algorithm described in (Cruz et al. 1998). For each pair of cyberlockers, we obtain a weight $t \in [0, 100]$. A weight of 100 indicates identical HTML structures; a value of 0 indicates entirely disjoint HTML structures. The rationale for this is to reveal if there are some individuals who control multiple cyberlockers by simply reusing the same or similar website templates.

Once the graphs are generated, we use the $Louvain$ algorithm (Blondel et al. 2008) to perform graph clustering. This is intended to identify communities based on shared characteristics in $Domains$, $Networks$ and $HTML$. In the case of $HTML$, this explicitly highlights cyberlockers that were likely generated by the same operator.

5.2 Understanding Cyberlockers

By exploring these three graphs an understanding of individual cyberlocker characteristics can be gathered, as well as providing insite into why and where these features are shared. Figures 5(a)(b)(c) present the graphs of *Domains*, *Networks* and *HTML*; the nodes within these graph are coloured to indicate the individual communities detected using the Louvain algorithm.

Shared Third-Party *Domains* Firstly, we look at the third party domains embedded within the cyberlockers. These include various domains ranging from ad networks to analytic services. In Figure 5(a) we identified 87 third-party domains (nodes) forming **4** communities. The globally most central nodes are three advertisement/tracking platforms — google.com (betweenness of 1,580), rtmark.net (536) and deloton.com (376). Structurally these create a fully connected graph with most cyberlockers embedding these domains (or being a maximum distance of 2). Employing the Ghostery database[8] we were able to classify 60% of extracted domains. Of these classified domains, 50% was classified as *First Party Exceptions*, 44% as *Advertising* and the rest split evenly between *Patterns*, *Site Analytics* and *Social-Media*.

As the predominant form of monetisation, we next explore the domains classified as *advertising*. Within these domains, popular ad networks include PopAds (in-degree 14), PubMatic (10) and DoubleClick (9). Note that PopAds specialises in "popunder" advertising (Le and Nguyen 2014) — something banned by Google's AdSense. Generally though, these major ad networks forbid publishers with illegal content and, therefore, their terms of service are clearly being broken (which risks account removal). The large number (23) of advertisement brokers used by these cyberlockers suggests that these policies are not strictly enforced. If these ad domains were to cease offering adverts to the cyberlockers the operators would likely be significantly affected. With this in mind, we see several alternative monetisation tools beginning to emerge. More than a third of the cyberlockers have started using the recently released *Coinhive* plugin which mines cypto-currencies on the viewer's machine. Furthermore, we observe the presence of various malicious domains *e.g. codeonclick.com* (4 cyberlockers), *rfihub.com* (3), and *nexac.com* (3), which download adware onto the viewer's machine, as well as *mathtag.com* (2), *exelator.com* (2) and *btrll.com* (2), which perform browser hijacking. The dependency that cyberlockers have on these revenue sources suggests their removal would pose a severe risk to their operation, unlike most prior P2P platforms which are self-sustaining.

Co-located *Network* Hosting Following on, we look at the Autonomous Systems (ASes) used to host video content. This is again modelled as a graph with links indicating the hosting of a particular cyberlocker within a given AS. This is important for several reasons. Most notably, the Digital Millennium Copyright Act (DMCA) allows legal parties to approach ASes who may have control over web content within their networks. These networks, therefore, represent a potential point of failure for the cyberlockers. In Figure 5(b), we see a total of **7** communities formed around such providers.

Firstly, we see three isolated communities: *thevideobee.com* (hosted on HISTATE), *veoh.com* (VEOH-AS) and watchers.to (PLI-AS). These jointly contribute only 3.4% of the streaming links from our selected set of cyberlockers. Thus, the impact of removing these websites (or ASes) would be limited. The remaining four communities are larger and inter-connected via a series of bridge nodes. Bridge nodes are, by definition, those websites that spread their content across *multiple* ASes. These are *openload.co*, *vidlox.tv*, *streamin.to*, *estream.to*, *vidzi.to*, and *vidto.me*).[9] Within the four inter-connected communities, M247 and Co-

[8] https://www.ghostery.com/

[9] Shutting down these streaming cyberlockers would remove 33% of all videos.

Figure 5: (a) Third-party domains linked by streaming cyberlocker homepages. (b) Autonomous Systems where video servers are hosted. (c) HTML similarity between cyberlocker homepages. Within each sub-figure, each colour indicates a community formed when the $Louvain$ method for community detection is applied. Note - some hosting ASes used by some sites are not present within (b) as these cyberlockers have not injected live videos during our crawl. In (b) and (c), the size of a node is proportional to its degree, while the thickness of an edge in (c) indicates the closeness in similarity

gent have the highest degree centrality, with a degree of 7 and 8 respectively (*i.e.,*they host 7 and 8 cyberlockers each). Shutting down these two ASes would result in 58% of the videos, and 71% of the servers observed in our set becoming unavailable. This indicates a remarkable level of vulnerability and clearly a point of failure that could be leveraged by copyright enforcers.

We posit that the 6 cyberlockers hosting content across multiple ASes may do so to increase resilience against takedowns (§6.2). Furthermore, to bolster this redundancy, there is a clear trend in the *physical locations* of the servers. M247 is based in Romania, which (as a country) hosts the largest share of streaming servers, containing 42% of the total streaming links witnessed. Similarly Cogent/Leaseweb are based in The Netherlands which hosts 23% of streaming links. This trend is reportedly driven by the lax copyright enforcement within these countries combined with their high capacity Internet infrastructure (Henderson 2013). A sudden increase in copyright regulation within these countries may see a shift in this behaviour and, again, we argue that this dependency on individual countries poses a resilience challenge for the cyberlockers.

***HTML* structure of cyberlocker homepage** Lastly, we compare the HTML of the cyberlocker homepages to detect underlying similarities between sites. This is because we hypothesise that some individuals may create *multiple* cyberlocker front-ends (*e.g.* to increase resilience). If this is true, it could imply that large segments of the cyberlocker system, which appear independent on the surface, are actually orchestrated by the same individual or organisation. To compute this, we use an existing pattern matching algorithm (Ratcliff and Metzener 1988), which gives each pair of websites a similarity score, $t \in [0, 100]$. To add context, we executed the algorithm on the Alexa Top 100 websites: the median similarity score was just 2.5. We then built a weighted graph with links between websites weighted by their similarity.

The resulting weighted similarity graph is presented in Figure 5(c). The thickness of an edge indicates the similarity score. However, unlike the previous bipartite graphs, the communities present show a *direct* (rather than transitive) relationship between cyberlockers. Within the graph two main communities can be identified, the Green group and the Yellow group. The median similarity scores in these two groups are 77.9 and 53.6 respectively. The Pink group contains the remaining very loosely connected cyberlockers with a median similarity of just 18.3. Manual inspection suggests that any scores above 30 indicate strong similarity.

It should also be noted that these similarities are also mirrored across $Domains$ and $Networks$. For example, the Green group all host within the Cogent AS. We further observe that 4 out of 6 sites (*gorillavid.com*, *gorillavid.in*, *movpod.net* and *movpod.in*) fall into the same community within $Domains$. These inferences are confirmed by WHOIS,[10] which reveals that *gorillavid*, *movpod*, *daclips* are all registered with the same owner. Unfortunately, 48% of our cyberlockers use WHOIS anonymisers to avoid registering their details. This prevents us from definitively proving shared ownership in other case. However, we briefly note that usage of anonymisers may itself indicate a similar owner. For instance, 0% of the sites in the Green group utilise anonymisers, in contrast to 73% in the Pink group.

It is also important to understand why these similarities emerge. In some cases, the similarities are driven by websites using similar templates. For example, *daclips*, *gorillavid* and *movpod* use the same structural template, but with different colour coding and other minor differences. The opposite extreme also exists — *novamov.com* operates

[10]A protocol for querying the registered owner of an Internet resource.

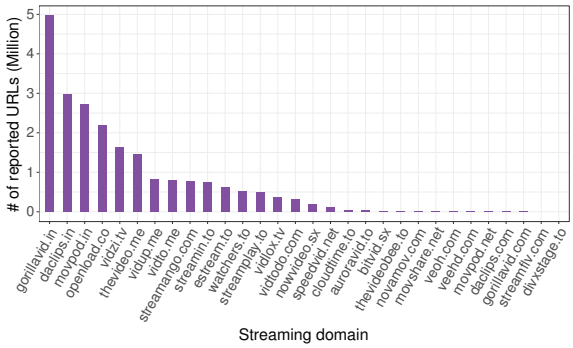

Figure 6: Number of video streaming URLs submitted as infringing belonging to streaming domains (Jan-Sep 2017).

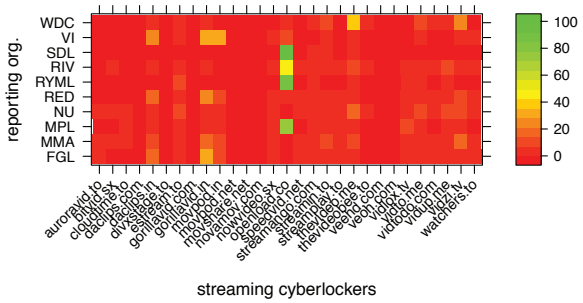

Figure 7: Percentage distribution of infringing URLs lodged by reporting organisations against each cyberlocker

as an front for *auroravid.to*; all visits to *novamov.com* redirect to *auroravid.to*. From the 28 node pairs identified as having a similarity scores above 75%, we find that 8 are identical pages and 8 have shared templates; the remaining 12 node pairs could be considered to share templates, but one or both of the nodes redirects to another page. From the identical pages, 3 are mirrored TLDs (*e.g.* .com and .in), whereas the remainder have totally different domains but pointing into a shared page. Overall, these observations mean that at least one fifth of our cyberlocker domains are actually operated by just two organisations/individuals, again confirming a remarkable dependency on just a small number of stakeholders.

6 Exploring Take down Dynamics

Finally, we briefly turn to our Lumen dataset to understand the level of copyright infringement taking place and the reactions of the websites to take down notices.

6.1 Overview of Complaints

In total we find 780M infringing URLs from our crawl of the Lumen database. This covers copyright complaints lodged from 01/01/17 to 30/09/17 and lists websites ranging from cyberlockers to Torrent sites. This list was, therefore, filtered to only include the 33 streaming domains observed in our crawl. This left 21.8M infringing URLs across 49,829 individual complaints from 304 organisations.

Figure 6 presents the number of complaints against each of the cyberlockers under-study. It can be seen that *gorillavid.in* has the most complaints by far, followed by *daclips.in* and *movpod.in* (note, these three were identified as existing in the same community in both the $Networks$ and $HTML$ graphs). These sites account for 48% of the total complaints made against our selected cyberlockers. Despite this, they do not constitute a major contributor to videos within our dataset (just 1.2%, 0.3% and 0.3% respectively). We see more contributions from their mirrored websites under different Top Level Domains (TLDs): *movpod.net*, *daclips.com* and *gorillavid.com* (4.7%, 4.5% and 3.8%). These mirrored sites received far fewer complaints, despite possessing more links. When combined, the top 3 most popular sites (*openload.co*, *thevideo.me* and *vidzi.tv*) receive 24% of all complaints.

Notice sender	No. of URL	No. of notice	No. of domains
fox group legal	8,508,289	1,876	22
markmonitor antipiracy	5,777,393	5,795	24
rivendell	3,369,745	9,842	26
vobile inc	3,214,057	12,044	28
redacted	400,329	228	28
remove your media llc	291,213	488	21
nbcuniversal	236,969	345	29
walt disney company	168,139	1,472	25
mg premium ltd.	97,176	1,063	26
skywalker digital ltd.	89,046	649	22

Table 2: Top 10 copyright infringing notice senders

This leads onto the question of who generates these complaints? We identify a total of 304 notice senders — Table 2 shows the top 10. Unsurprisingly, these are dominated by content producers such as 21st Century Fox and Walt Disney. We also find a number of dedicated anti-piracy companies (*e.g.* Mark Monitor, Rivendell, Vobile). These top 10 notice senders contribute 96% of all URLs complained about within the list examined, with the remaining 294 covering just 4%. Upon closer inspection, trends can be observed among these top complainants. Figure 7 presents a heat map; the Y-axis list the top 10 complainants, the X-axis lists the cyberlockers. The heat represents the fraction of notices from each complainant to each cyberlocker. It can be seen that most complainants are highly selective in terms of which cyberlockers they complain about. For example, about 98%, 89% and 73% of URLs complained about by *Skywalker Digital*, *Remove Your Media* and *MG Premium Limited* were aimed at *openload.co*. Why such organisations choose to target individual cyberlockers in this way is unclear. However, the trend generalises across other complainants too. For instance, we see *daclips.in*, *gorillavid.in* and *movpod.in* being jointly targeted by *Fox Group* (63%), *Mark Monitor* (38%), *Redacted* (55%) and *Vobile* (86%). Despite this, we find no evidence that these cyberlockers contain more or less content belonging to each complainant.

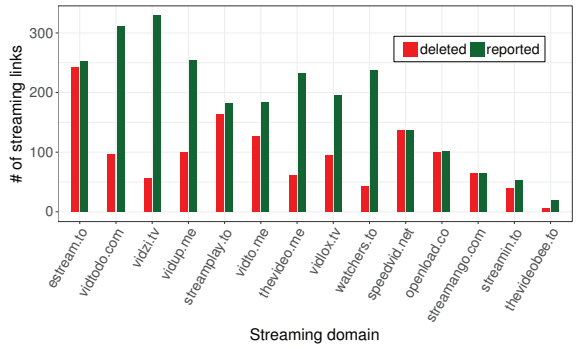

Figure 8: Cyberlocker URLs reported for infringing copyright compared to URLs deleted. X-axis ranked by fraction of takedown requests acted upon.

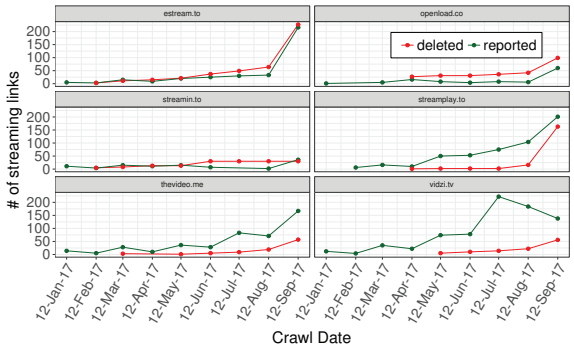

Figure 9: Time-series of removals. A reported link is available if at the month of crawl we can access the video.

6.2 How Do Cyberlockers React?

The above shows that complaints are regularly made against these portals. Next we inspect the reaction of cyberlockers to such complaints. Note, Lumen does not record specific complaints made *to* the cyberlockers, they record complaints made *about* them (to other parties *e.g.* Google, Bing). Utilising the monthly snapshots, we can see if videos uploaded in 2017 were removed after a complaint had been lodged. Thus, we extract the set of complaints that correspond to videos in our dataset. This leaves a total of 2,669 streaming links reported in 2017, associated with 275 videos released during this period. This, of course, leaves a large number of complaints in our Lumen database that we do not have corresponding video data for. This, unfortunately, is inevitable due to the sheer scale of the ecosystem.

Figure 8 plots the number of complaints and the number of removals for each cyberlocker across our entire measurement period. Only cyberlockers which received requests to remove links gathered during our monthly crawls are included in the figure. Within the figure, if all videos complained about were removed, then the number of removals and the number of complaints would be equal. The X-axis is ordered by the fraction of complaints acted upon. A clear ranking can be seen with some cyberlockers removing nearly all videos complained about[11] (*e.g. estream.to*), whereas others (*e.g. vidzi.tv*) ignore nearly all complaints. We can compare this "obedience" rank against the others previously discussed. We find little correlation between this and the Alexa rank (Spearman Rank -0.015), but a stronger correlation with the number of links on the site (-0.81). This might exists because larger sites find it more difficult to ignore legal pursuit.

To expand on this we can also explore the removals over time. Due to space constraints we select 6 cyberlockers that have a mixture of behaviours. For these we plot the number of videos reported and deleted over time in Figure 9. We observe a variety of behaviours. For example, websites such as *openload.co*, *estream.to* and *streamin.to* react positively to copyright reports: over 75% of videos are deleted within 1 month of complaints being registered on Lumen. The same cannot be said of *vidzi.tv* and *thevideo.me*, where <30% of videos are deleted within 1 month. We observe that the videos that are *not* deleted from *openload.co*, *estream.to*, *vidzi.tv* are all hosted in Romania on M247 (the videos that are deleted are in other ASes). That said, it would be unwise to draw conclusions here, as Romania hosts both the cyberlocker that ignores the most complaints *and* the cyberlocker that acts upon most complaints. Overall, the country hosting content that least frequently respects complaints is the Netherlands, where only 6% of requests are acted upon. Hence, the diversity seen within individual countries suggests that the decision to act upon a complaint is largely driven by the individual cyberlockers.

[11] Note we cannot conclude that the removal was directly caused by the takedown notice. We can only assume this is likely.

7 Conclusion & Future Work

In this paper, we have offered a first study of the emerging streaming cyberlocker ecosystem. We began by exploring the streaming links shared on indexing sites (§4). We discovered an environment actively involved in copyright infringement with an aggressive injection of recent releases. We proceeded to examine the individual characteristics and potential relationships between these websites (§5). This identified a number of communities based on shared domains, shared hosting facilities and high levels of HTML similarity. In some cases we found that this was individual operators running multiple cyberlocker instance or simply redirecting into the same (or mirrored) websites. This may be done for many reasons, but we believe it is most likely to increase resilience in the face of legal action (§6). Indeed, a common observation is the vulnerability of the cyberlockers. For instance, we find that over half of all content observed is hosted within just two ASes.

This is just the first step in our exploration of the cyberlocker ecosystem. There are a number of future lines of work we will explore. We emphasise that our data has only inspected a *slice* of the streaming cyberlocker ecosystem. There are many other indexing sites, as well as portals that we have not investigated yet. This is evidenced by the number of complaints on Lumen that we did not have the corresponding video data for. Hence, our major line of future

work is expanding our datasets to generalise findings across a broader swathe of the ecosystem. We also wish to further investigate the *relationships* between cyberlockers. This will involve building more graph-based models, underpinned by alternative data such as feature extraction from HTML. We further plan to investigate longitudinal trends. For example, we believe websites may periodically "rebrand" after they have previously been taken down — studying this as an evolving series of graphs will help identify this. We also wish to better understand the evasion tactics used by these sites, particularly in the face of changing strategies used by copyright enforcers. This will, of course, involve deep diving into the way that enforcers select websites to complain about.

References

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. 1–12.

Cha, M.; Kwak, H.; Rodriguez, P.; Ahn, Y.; and Moon, S. 2007. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. *ACM Internet Measurement Conference (IMC)* 1–14.

Clay, A. 2011. Blocking, tracking, and monetizing: Youtube copyright control and the downfall parody. Institute of Network Cultures: Amsterdam. 219–233.

Cruz, I. F.; Borisov, S.; Marks, M. A.; and Webb, T. R. 1998. Measuring structural similarity among web documents: preliminary results. In *Electronic Publishing, Artistic Imaging, and Digital Typography*. Springer. 513–524.

Ding, Y.; Du, Y.; Hu, Y.; Liu, Z.; and Wang, L. 2011. Broadcast yourself: understanding YouTube uploaders. *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference* 361–370.

Dutta, R., and Patel, K. C. 2008. Detecting copyright violation via streamed extraction and signature analysis in a method, system and program.

Elder, R. 2016. Illegal streaming is dominating online piracy.

Farahbakhsh, R.; Cuevas, A.; Cuevas, R.; Rejaie, R.; Kryczka, M.; Gonzalez, R.; and Crespi, N. 2013. Investigating the reaction of bittorrent content publishers to antipiracy actions. In *13th IEEE International Conference on Peer-to-Peer Computing*, 1–10.

Henderson, A. 2013. The five best countries to host your website for data privacy and fastest internet speed in the world.

Hilderbrand, L. 2007. Youtube: Where cultural memory and copyright converge. *FILM QUART* 61(1):48–57.

Lauinger, T.; Onarlioglu, K.; Chaabane, A.; Kirda, E.; Robertson, W.; and Kaafar, M. A. 2013a. Holiday pictures or blockbuster movies? Insights into copyright infringement in user uploads to one-click file hosters. *16th International Symposium on RAID* 369–389.

Lauinger, T.; Szydlowski, M.; Onarlioglu, K.; Wondracek, G.; Kirda, E.; and Kruegel, C. 2013b. Clickonomics: Determining the Effect of Anti-Piracy Measures for One-Click Hosting. *Network and Distributed System Security Symposium* 1–14.

Le, T. D., and Nguyen, B.-T. H. 2014. Attitudes toward mobile advertising: A study of mobile web display and mobile app display advertising. *Asian Academy of Management Journal* 19(2):87–103.

Mahanti, A.; Carlsson, N.; Arlitt, M.; and Williamson, C. 2012. Characterizing cyberlocker traffic flows. *Proceedings - Conference on Local Computer Networks* 410–418.

Miaoran, L. 2009. The pirate party and the pirate bay: How the pirate bay influences sweden and international copyright relations. *Pace Int'l L. Rev.* 21:281.

Monitor, M. 2011. Traffic report: Online piracy and counterfeiting.

NetNames. 2014. Behind the Cyberlocker door : Use Credit Card Companies to Make Millions.

Orlowski, A. 2013. Brit isps ordered to add more movie-streaming websites to block list-the register.

Rafique, M.; Goethem, T. V.; Joosen, W.; Huygens, C.; and Nikiforakis, N. 2016. It's Free for a Reason: Exploring the Ecosystem of Free Live Streaming Services. *23th Annual Network & Distributed System Security Symposium* 21–24.

Ratcliff, J. W., and Metzener, D. E. 1988. Pattern matching: the gestalt approach. *Dictionary of Algorithms and Data Structures*.

Raustiala, K., and Sprigman, C. 2012. How much do music and movie piracy really hurt the u.s. economy?

Sanjuàs-Cuxart, J.; Barlet-Ros, P.; and Solé-Pareta, J. 2012. Measurement based analysis of one-click file hosting services. *Journal of Network and Systems Management* 276–301.

Sastry, N. R. 2012. How to tell head from tail in user-generated content corpora. In *International AAAI Conference on Weblogs and Social Media*.

Torres, R.; Finamore, A.; Kim, J. R.; Mellia, M.; Munafò, M. M.; and Rao, S. 2011. Dissecting video server selection strategies in the YouTube CDN. *Proceedings - International Conference on Distributed Computing Systems* 248–257.

Zink, M.; Suh, K.; Gu, Y.; and Kurose, J. 2008. Watch Global, Cache Local: YouTube Network Traffic at a Campus Network - Measurements and Implications. *Proceeding of the 15th SPIEACM Multimedia Computing and Networking (MMCN)* 6818:28.