Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

MagnetDB: A Longitudinal Torrent Discovery Dataset with IMDb-Matched Movies and TV Shows

Scott Seidenberger, Noah Pursell, Anindya Maiti

University of Oklahoma
seidenberger@ou.edu, noah.a.pursell-1@ou.edu, am@ou.edu

Abstract

BitTorrent remains a prominent channel for illicit distribution of copyrighted material, yet the supply side of such content remains understudied. We introduce *MagnetDB*, a longitudinal dataset of torrents discovered through the BitTorrent DHT between 2018 and 2024, containing more than 28.6 million torrents and metadata of more than 950 million files. While our primary focus is on enabling research based on supply of pirated movies and TV shows, the dataset also encompasses other legitimate and illegitimate torrents. By applying IMDb-matching and annotation to movie and TV show torrents, *MagnetDB* facilitates detailed analyses of pirated content evolution in the BitTorrent network. Researchers can leverage *MagnetDB* to examine distribution trends, subcultural practices, and the gift economy within piracy ecosystems. Through its scale and temporal scope, *MagnetDB* presents a unique opportunity for investigating the broader dynamics of BitTorrent and advancing empirical knowledge on digital piracy.

Introduction

Digital piracy has long posed a challenge to content publishers, policymakers, and internet service providers around the globe (US Chamber of Commerce 2019; Danaher, Smith, and Telang 2020). While digital piracy can manifest through numerous channels, BitTorrent (BitTorrent 2025; Pouwelse et al. 2005) stands out as one of the most prevalent networks used to facilitate unauthorized peer-to-peer (P2P) sharing of music, films, software, and other media (Digital Music News 2024; Lifewire 2024a; Ancell 2024). Movies and television shows in particular dominate a significant portion of the pirated content (Wang, Liu, and Xu 2010), reflecting the ongoing supply and demand for entertainment multimedia. Streaming service fragmentation, where exclusive content is spread across multiple platforms, has recently led to increased piracy as consumers become frustrated with managing and affording numerous subscriptions (Ellis 2023; Bode 2018; Lifewire 2024b). BitTorrent's decentralized architecture, ease of use, and recent innovations like streaming capabilities through WebTorrent (WebTorrent 2025) sustain its role as a prominent platform for unauthorized distribution, creating a fascinating sociotechnical ecosystem for studying its operation and evolution over time.

In this work, we introduce *MagnetDB*, a large torrent database continuously capturing long-term, real-world BitTorrent activity with explicit focus on the supply side of pirated media. Our motivation aligns with the broader social media and web research communities, which can benefit from examining the cultural and social dynamics behind media consumption, distribution, and public engagement with unauthorized content. The potential applications of *MagnetDB* are numerous. Cultural analytics researchers can study the availability and popularity of movies or TV shows and how these patterns relate to production timelines or cultural events. Linguists and anthropologists may look at naming conventions and tagging practices within torrent metadata, examining how language usage reflects subcultural identities across different release groups and communities. For policymakers and industry practitioners, the dataset can inform anti-piracy strategies, guide legal service offerings, or provide a more accurate picture of the scope and scale of unlicensed distribution. With a focus on the supply side, *MagnetDB* also enables unique investigations into the motivations and behaviors of torrent creators (encoders), shedding light on "gift economy" dynamics within piracy ecosystems (Huizing and van der Wal 2014).

MagnetDB spans more than five years of torrent discovery (December 2018 – September 2024), containing **28,606,694 torrents** with **950,660,089 files**, which total **82.87 petabytes** of shared media. Although *MagnetDB* contains torrents across the broader BitTorrent ecosystem, our primary focus is to enable research based on the supply of pirated movies and TV shows. To achieve this, we additionally employ a metadata matching process and identified **1,562,573** movie and TV show video files covering **78,740** unique titles on IMDb (IMDb 2024), and annotated those files with corresponding IMDb identifiers within *MagnetDB*. By cross-referencing pirated video files with rich IMDb metadata, such as genre, release year, ratings, cast, runtime, and reviews, *MagnetDB* enables more granular analyses of distribution patterns, popularity trajectories, and usage behaviors that surpass what can be inferred from file-level information alone.

While a significant portion of BitTorrent usage involves unauthorized file sharing, it is also important to recognize

that many torrents facilitate legitimate uses, including open-source software releases, academic datasets, and other legal media distributions. Nonetheless, by encompassing both pirated and legitimate torrents, *MagnetDB* enables a more comprehensive exploration of file-sharing practices, usage trends, and broader distribution dynamics in the BitTorrent network.

The Scene

Often referred to as "The Scene," this loosely knit yet highly organized network of groups and individuals has played a pivotal role in the distribution of pirated content since the days of dial-up bulletin board systems. Historically, The Scene has been most closely identified with the so-called "warez" community, which initially focused on cracking and distributing software. Over time, their activities expanded to music, films, television episodes, e-books, and other digital media. Although The Scene's structure can appear disjointed from the outside, most participants adhere to an implicit code of conduct and a federated organizational structure designed to promote a sense of community among its affiliates and produce quality warez (Goldman 2005; McCandless 1997).

Key Players and Motivations. Scene groups typically occupy specialized roles in the supply chain: "release groups" focus on sourcing and preparing new content (e.g., by obtaining pre-release copies of software or media), while "couriers" and other intermediaries distribute the files across private platforms, racing to achieve "zero-day" status, making content available on or before the official commercial release. In this environment, prestige is closely tied to speed: the faster a group can acquire and release a highly anticipated title, the more its reputation grows within the community (Goldman 2005).

As *MagnetDB* provides comprehensive data on the torrent supply, it can provide significant insight into the output of the release groups. It is important to note some key taxonomic definitions, as there is an important distinction within the release groups. The groups or individuals that *create* the torrent are the "encoders", and the groups that *distribute* the torrent information are the "sites" (websites). *MagnetDB* makes a distinction between these two entities in the dataset.

This subculture operates largely as a gift economy, where members thrive on recognition and status rather than monetary gain (Hétu, Morselli, and Leman-Langlois 2012; Rehn 2004). For many participants, the act of collecting and sharing files is itself the reward, having a full library of releases is a point of pride, irrespective of whether those files are ever personally used or consumed.

Cultural Underpinnings: Gift Economy and Reputation. Unlike purely profit-driven piracy rings, The Scene is upheld by a tradition of reciprocity and social capital accumulation. Members who consistently supply high-quality or hard-to-find releases gain respect and clout, which can lead to better access to private servers, exclusive pre-releases, and membership in elite groups. This gift economy ethos is driven by a combination of altruistic sharing norms, social bonding, and competitive one-upmanship. In essence, Scene participants work to cultivate an identity not just as savvy digital sharers, but also as gatekeepers who uphold community standards—particularly regarding file quality, completeness, and timeliness. Notably, Huizing and van der Wal observe that these cultural imperatives can wax or wane over time, as once-lively sub-communities (e.g., the MP3 Scene) either mature, splinter, or face changing technological and social pressures (Huizing and van der Wal 2014; Danaher, Smith, and Telang 2020).

Opposition and Enforcement Challenges. On the other side of The Scene's shadow economy stand content owners, such film studios, record labels, and gaming companies. They have formed industry consortia who have marshalled significant legal and technological forces to clamp down on unauthorized distribution. Organizations such as the Motion Picture Association (MPA) and the Recording Industry Association of America (RIAA), as well as various government and law enforcement agencies, actively track and pursue major release groups. High-profile investigations have led to global raids, arrests, and lawsuits, but these crackdowns often disrupt only portions of The Scene's vast and decentralized network. Owing to its resilience and the strong incentives for members to maintain secrecy, new sites, servers, and courier collectives frequently emerge even as others are shut down (Basamanowicz and Bouchard 2011).

Contemporary Developments and Ongoing Tensions. Despite the persistent threat of legal action, The Scene continues to exert a formative influence on wider piracy practices, including public torrent trackers. Advances in encryption, decentralized storage, and VPN services have enabled the continuation of these piracy practices despite ever-intensifying enforcement. Newer decentralized P2P networks, such as IPFS, are being used to host and distribute pirated content (Shi et al. 2024).

As David (David 2017) argues, free-sharing networks operate at near zero-marginal cost, thereby subverting the core scarcity principle of market-based economies and perpetuating what he terms "a crime against capitalism." This tension between decentralized file-sharing, premised on unconstrained replication, and industry-backed intellectual property regimes underscores the ongoing challenges of enforcing exclusivity in a post-scarcity environment.

The Scene exemplifies the interplay between technological innovation, cultural norms, and external pressures that shape file-sharing ecosystems. Its participants cultivate a reputation-based hierarchy that rewards efficiency, while those seeking to control or criminalize their activities face the practical challenge of shutting down a network whose primary currency is prestige rather than profit. As such, The Scene's influence remains visible not only in the rapid release of popular media but also in the evolving legal, technical, and social landscapes surrounding file sharing more broadly.

Background and Related Work

BitTorrent and the DHT. A basic unit of distribution on BitTorrent is the *torrent*, an informational file or magnet link containing metadata such as filenames, sizes, and cryptographic hashes. Tying together *swarms* of peers, each tor-

rent spreads through a cooperative process wherein *seeders* (users possessing the complete file) and *leechers* (users still downloading) exchange pieces of the shared file. Although many trackers remain popular for indexing torrents (The Pirate Bay 2003; 1337x 2007; RARBG 2008; LimeTorrents 2009), including historic torrents and torrents without any network activity, they often have limited coverage of real-world torrent activity as they depend on the encoder to add trackers to their torrents. To address this limitation and obtain a real-world view of actively shared content, we utilize the BitTorrent Distributed Hash Table (DHT) to discover new torrents directly from the P2P network, ensuring that our dataset captures a broader and more dynamic range of content as it emerges. *While BitTorrent DHT enhances resiliency and scale, it also complicates attempts to study the network comprehensively, requiring torrent discovery from a diverse set of peers (which utilizes significant bandwidth and compute) and over extended periods.* BTDigg (BTDigg 2025), a search engine for torrents discovered on the DHT, provides a useful means to explore decentralized torrent activity, though its dataset is not readily available for direct analysis.

Analysis of Peer Activity and Torrents. Despite prior research on piracy in the BitTorrent network, our understanding of the *supply side* of pirated content remains limited. Existing studies focus on tracking seeders and leechers (Zhang et al. 2011; Le Blond et al. 2010) or on the DHT infrastructure as it tracks P2P swarm behaviors (Wang and Kangasharju 2013). However, these works often lack comprehensive content metadata and do not fully capture the complexities of what exactly is being shared and by whom. Prior works that do analyze content metadata (Wolchok and Halderman 2010) typically fail to provide a longitudinal scope for observing shifts in piracy trends over several months and years. Consequently, there is a need for a large-scale dataset integrating long-term observation with rich content metadata in order to advance the empirical understanding of piracy factors in the BitTorrent network.

Datasets with Torrent Metadata. Only a few existing datasets provide a long-term, consistent view of torrent discovery. The Torrent Metadata Archive hosted on Archive.org (Internet Archive 2023) contains metadata for 83 million torrents discovered between 2016 and 2023, though details about its collection process and consistency are unclear. The Kiwi Torrent Research dataset (Kiwi Torrent Research 2023b,a) combines torrents from three different sources, including the Archive.org dataset and an earlier checkpoint of our *MagnetDB* dataset, highlighting *MagnetDB* as a recognized and significant contribution to the research community. As comprehensive and well-documented torrent datasets remain scarce, *MagnetDB* stands out as an ongoing effort emphasizing consistency in torrent discovery.

Dataset and Collection Methodology

Torrent Data

To discover torrents and record their metadata, we deployed the open-source BitTorrent DHT crawler `magnetico` on our self-hosted vantage point (Alper 2020). This software continuously traverses the decentralized BitTorrent DHT, iteratively seeking out nodes in the network and aggregating torrents discovered from each node. As new nodes are learned, the crawler systematically explores their advertised torrents, building a comprehensive index of accessible swarm information. To enhance the crawler's reach and ensure a broader exploration of the DHT, our `magnetico` instance is configured with a customizable parameter *indexer-max-neighbors=10000*, enabling it to track and index torrents from up to 10,000 peers simultaneously at any given time.

Over the 300-week collection period—from December 30, 2018, through September 29, 2024—we gathered **28.6 million torrents** consisting of more than **950.6 million individual files**. Table 2 summarizes our final collection counts and the subsequent processing steps. Throughout this time, our crawler operated at an **overall uptime of 93.7%**, with occasional lapses due to hardware maintenance and network outages. When the system is down, it did not collect new data (shown in red in Figure 1), while degraded periods (shown in orange) reflected reduced coverage, defined here as weeks where fewer than 10,000 new torrents were discovered.

Figure 1 illustrates how the incoming torrent discovery rate initially spiked at the outset of our deployment, reflecting the so-called "burn-in" period. During burn-in, `magnetico` rapidly indexes pre-existing torrents advertised by the DHT. Afterward, discovery stabilizes around a more consistent, long-term rate, capturing newly published torrents as they appear. Research on BitTorrent suggests that the network behaves closest to a random graph, rather than a scale-free or small-world network (Su et al. 2013). However, the network still exhibits a small diameter (< 6), which helps in the propagation of new information especially on the DHT. Thus, while we consistently observe that most new torrents appear in our database shortly after their initial release, some swarms remain difficult to reach if they have relatively few or transient peers. Manual verification of *MagnetDB* against known publication times on major torrent distribution sites (such as YTS) confirms that new torrents typically appear in our database on the same day they are released.

Notably, operating a resource-intensive system of this scale demands substantial internet bandwidth, approximately **30 TB per month**, with a roughly 2:1 ratio of downloaded to uploaded data. While this ensures deeper coverage of the DHT, researchers aiming to replicate or extend this work should carefully consider both the computational cost and the network-level impacts of large-scale crawling.

File Processing

Each torrent can have several files associated with them. In the case of multimedia, the files can be a combination of media files and non-media files. Non-media files can include a text-based "credit" file with the information of the uploader, subtitles, or metadata files about the media files. For TV shows, there may be several episodes each as a file in a single torrent. For movies, the movie is typically a single file in the torrent, but in some cases there may be different ver-

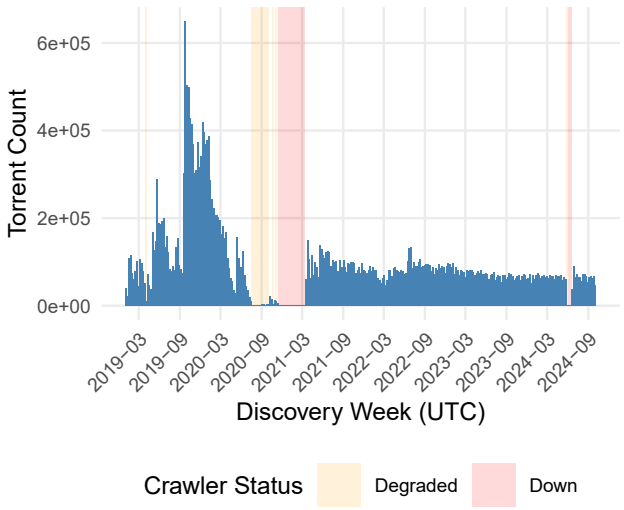

Figure 1: Histogram of discovered torrents with crawler downtime (red) and degraded periods (orange). Degraded is defined as weeks where less than 10,000 torrents were discovered. Uptime over 5 years and 9 months was 93.7%.

sions of the same movie in the same torrent, or a sample of the movie as a separate file. Since we focus on multimedia in the form of movies and TV shows, we first identify whether each file is a video file by its file extension.

As discussed earlier in our section on *The Scene*, suppliers of the bulk of multimedia torrents are known to be relatively well-organized groups, collectively affiliated with The Scene subculture. This subculture voluntarily abides by strict conventions in the creation and naming of pirated content (Basamanowicz and Bouchard 2011; Dementiev and Fenton 2016). A democratic, small council of the top suppliers of pirated content regularly publish the accepted standards that groups follow, and we find consistency with these standards in our dataset. We use these standards to parse the torrent files for the title as well as all the metadata fields.

> **Movie**
> `Feature.Title.<YEAR>.<TAGS>.[LANGUAGE]`
> `.<RESOLUTION>.<FORMAT>-GROUP`
>
> **TV Show**
> `Weekly.TV.Show.[COUNTRY_CODE].[YEAR]`
> `.SXXEXX[Episode.Part].[Episode.Title]`
> `.<TAGS>.[LANGUAGE].<RESOLUTION>.<FORMAT>-GROUP`
>
> https://scenerules.org/

Using a library that leverages these patterns (Bindlish 2016), we attempt to parse from each torrent file's name as many metadata fields as possible. While Table 1 shows examples of the metadata fields extracted, the complete list can be found in the documentation `README` with *MagnetDB*.

Content	Distribution	Technical
documentary	site	audio
3d	website	codec
genre	network	resolution
language	repack	widescreen
directorsCut	readnfo	hdr
internationalCut	internal	fps
unrated	region	bitDepth
remastered	encoder	fps

Table 1: Example of types of metadata extracted from file names and torrents.

Title Matching and Metadata Extraction

The next step in the analysis of these multimedia files is to match them against the IMDb dataset (IMDb 2024). Accurate matching allows us to verify the release dates of the content and narrow-down our area of interest to movies and TV shows. We populated an Elasticsearch index with the IMDb dataset. The index was configured with a custom analyzer designed to optimize the matching process for our specific use case.

The custom analyzer employs an edge n-gram filter, which generates n-grams from the beginning of a word up to a specified length. Specifically, we set the minimum n-gram length to 4 and the maximum to 15. This configuration captures meaningful fragments of words while avoiding overly short or excessively long n-grams that might lead to false positives or negatives. The analyzer also incorporates standard text processing steps such as lowercasing and ASCII folding, normalizing the text by converting it to lowercase and replacing accented characters with their ASCII equivalents. This normalization ensures that variations in capitalization or the presence of special characters do not hinder the matching process.

For each title, we construct a search query to retrieve the most similar entries from the IMDb index. Elasticsearch's default scoring algorithm, BM25 (Best Match 25) (Robertson, Zaragoza et al. 2009), is used to rank the search results based on their relevance to the query. The BM25 algorithm calculates a relevance score by considering factors such as term frequency, inverse document frequency, and document length normalization. This scoring method effectively ranks potential matches by their textual similarity to the query, which is essential for accurately matching titles that may have slight variations or idiosyncrasies. BM25 scores are positive numbers that can vary widely depending on the query and the dataset and are unbounded. A higher BM25 score indicates a higher relevance between the query and the document. While the choice of threshold is subjective based on the objective of the analysis, the results of the specific implementation of BM25 do not practically impact search effectiveness, with a large scale reproducibility study confirming that BM25 implementations do not yield significant differences between them (Kamphuis et al. 2020).

After retrieving the top matches from the Elasticsearch index, we evaluate the relevance scores to determine if a suitable match has been found. If the highest-ranked IMDb en-

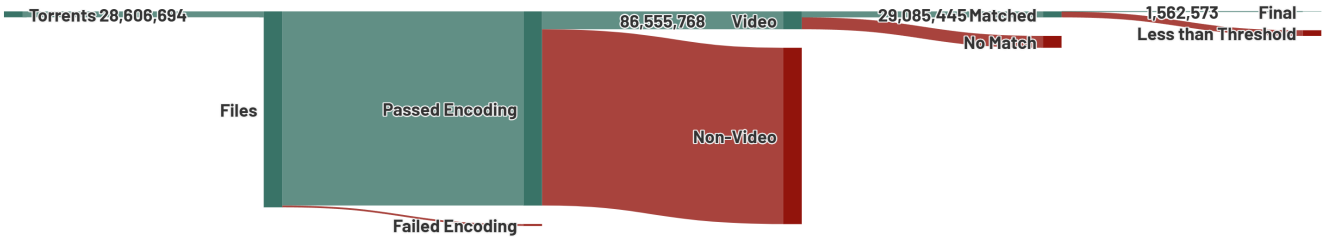

Figure 2: Sankey diagram of the data processing pipeline going from torrents to the final matched dataset. *MagnetDB* provides all data at each processing step so that other researchers can tailor their own data pipeline to fit their use case.

try has a score exceeding a predefined threshold, we consider it a match and associate the corresponding IMDb metadata with the torrent file. This metadata includes the official title, release year, genre, and other details that enrich our dataset. For the purpose of our initial exploration, we chose a 2σ threshold match score of 138, illustrated in Figure 4, retaining 1.81% of the video files, reducing the likelihood of false positives—cases where a torrent title is incorrectly matched to an unrelated IMDb entry due to generic or common terms. For an example of matches with high and low scores, see the example in the Appendix. The entire processing pipeline is illustrated in Figure 2 as a Sankey diagram. It is important to note that this data pipeline can be modified to fit any use case, as we provide all the original data.

Descriptive Statistics

In this section, we provide a broad overview of *MagnetDB*, detailing the scope and composition of the dataset, the characteristics of the files and torrents it contains, as well as the subset of videos successfully matched to external metadata resources like IMDb. Table 2 summarizes our key statistics, and the following subsections expand on these figures and the accompanying visualizations.

Statistic	Count
# of torrents	**28,606,694**
# of total files (passed encoding)	**942,076,233**
# of files with encoding failures	8,583,857
# of video files	86,555,768
# of non-video files	855,520,465
# of matched files (IMDb)	29,085,445
of which, **final (above threshold)**	**1,562,573**
# of matched but below threshold	27,522,872
Match rate (% of video files matched)	33.6
Final match rate (% of video files, above threshold)	1.81
# of movies	$\approx$ 751,256
# of TV episodes	$\approx$ 811,316
Avg. # of files per video torrent	$\approx$ 3.026

Table 2: Final counts and statistics.

Torrents (Swarms)

Each torrent, or *swarm*, represents a distinct set of peers simultaneously sharing or downloading a collection of files. Our collection spans **28.6 million** torrents, containing over **942 million** individual files which passed encoding checks. **86.56 million** ($\sim$9.2%) are identified as video files based on their extensions. As shown in Table 2, approximately 3.0% of all files were discarded during processing due to encoding failures. Across all torrents, we observe a mean of 33.23 files per torrent, illustrating the multi-file structure commonly employed—particularly for large multimedia releases that bundle multiple episodes or file versions.

Should one attempt to download every torrent in *MagnetDB*, the combined size would exceed **82.87 PB**. Though the vast majority of torrents include only tens or hundreds of megabytes of data, a few contain massive aggregates of multi-terabyte content, such as archival collections.

Files

Figure 3 shows the distribution of the file sizes on a log scale. The histogram portion depicts the density, whereas the line represents the cumulative distribution function (CDF). Most files fall between 1 KB and 1 GB, reflecting typical document, audio, and video content. A non-negligible fraction (over 170,000 files) exceeds 1 GB, attesting to the prevalence of high-definition media content—especially movies and TV episodes, which can range from a few hundred megabytes to tens of gigabytes each. An extreme upper tail (40 files in the range of 1 TB or more) largely corresponds to large archival releases or entire libraries distributed in a single torrent.

Videos Matched to IMDb

Of the 86.56 million video files, around **29.09 million** ($\sim$33.6%) were potential matches to IMDb based on the title matching procedure described in the *Title Matching and Metadata Extraction* section. However, to ensure sufficient confidence and minimize false positives, as illustrated in Figure 4, we impose a 2σ BM25 score threshold of 138 (the red dashed line). This yields a final matched set of **1.56 million** video files, or $\sim$1.8% of all videos. These files collectively amount to 1.85 PB of data–just over 2% of the

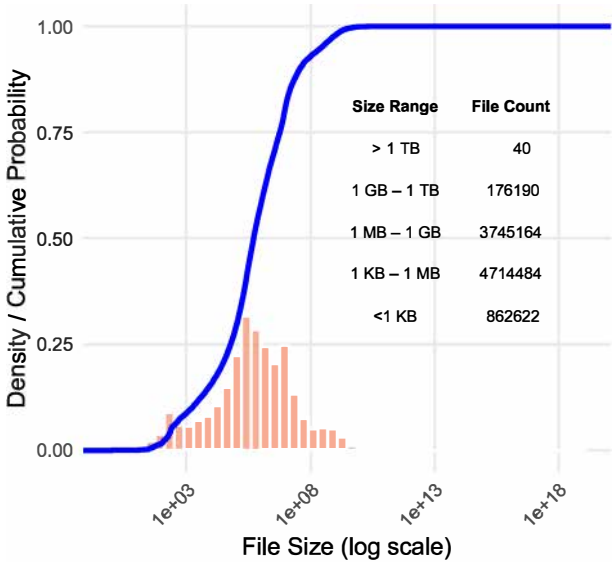

Figure 3: Density plot with overlaid CDF of file sizes in bytes.

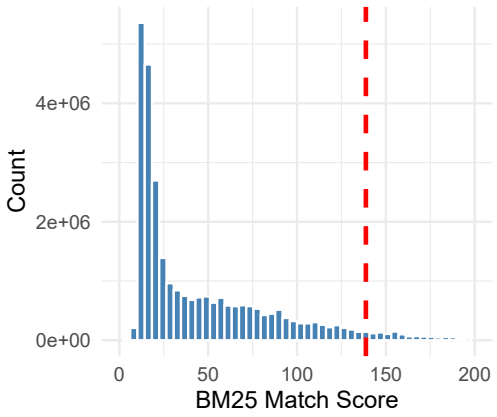

Figure 4: Distribution of BM25 Match Scores to the IMDb Non-Commercial Database. The Red-dashed Line indicates the 2σ (138) threshold past which we retain rows for the IMDb-matched subset.

dataset's total size. Within the final matched dataset, we estimate 751K distinct movies and 811K TV episodes.

Figure 5 illustrates breakdowns of some of the metadata fields in the matched dataset by video *quality* (e.g., WEB-DL, BluRay), *language* (e.g., English, Russian, Polish), *site* (the groups/websites that distribute the torrent files), and *encoder* (the group or individual responsible for preparing the media). While English and Russian stand out among labeled languages, there is a long tail of other languages. Similarly, a few large release sites and encoders dominate, but countless small or independent groups also contribute to the BitTorrent landscape.

Targeting of Streaming Services. Figure 6 shows the breakdown of identified titles in *MagnetDB* by their affiliated streaming service or production company. Notably, large commercial platforms dominate the high end of the spectrum, with *Amazon MGM Studios* leading, followed by *Netflix* and the *BBC*. *Disney Plus*, *Hulu*, and *HBO Max* also occupy substantial niches, illustrating that a range of well-funded, global streaming services face high targeting rates within the BitTorrent ecosystem. These findings underscore the continuing pull of premium digital media in the piracy community, wherein the most prominent and prolific services are also the most widely pirated.

From a supply-side perspective, this aligns with the general pattern that release groups and encoders concentrate on popular or high-demand works. Commercial platforms like Amazon and Netflix invest heavily in exclusive content—often releasing entire seasons of shows at once—which encourages immediate torrenting. Meanwhile, the BBC's position in the top tier attests to worldwide interest in British-produced series and documentaries, some of which remain difficult to access legally outside the United Kingdom. In essence, these results hint at the interplay between streaming exclusivity, global distribution gaps, and user demand. By examining which platforms rank highest in pirated titles, *MagnetDB* offers valuable insights into how audience preferences and licensing strategies translate into increased torrent availability.

Coverage Over Time. A central question when evaluating torrent datasets, or any piracy-oriented data, is the degree of coverage in terms of movies and TV shows against a well-known reference, here IMDb. Figure 7 shows how the matched dataset of *MagnetDB* intersects with IMDb entries over time. The histogram portion indicates the total counts of titles in *MagnetDB* by release year, and the line shows the percentage coverage relative to IMDb's total corpus for that year. The coverage is below 5% for most years, showing that there is far more content that has been created than is actively torrented.

Interestingly, a notable spike occurs in the early-to-mid 1940s. This unexpectedly high coverage in the wartime era likely reflects the enduring popularity of certain classic titles (e.g., *Casablanca (1942)* and *Citizen Kane (1941)*) coupled with the cultural significance of World War II films, which remain actively shared. Additionally, it was in the 1940s that Paramount introduced a new projection system that made color production more affordable.

For more contemporary releases (post-2000), there is a dramatic surge in IMDb's catalog, fueled by both increasing global film production and the proliferation of digital-native content such as web series and independent online releases. By contrast, the community of encoders responsible for torrents has not scaled proportionately, and only a fraction of these newer titles appear in *MagnetDB*. This suggests a practical bottleneck on what gets shared via BitTorrent. Scene groups and independent encoders often focus on popular or high-demand works, leaving a large swath of minor or niche titles unrepresented. As IMDb continues to grow–reportedly adding over 10,000 new titles per day–this gap may widen. Future research could employ *MagnetDB* to study *why* certain works gain traction in the torrent ecosystem while oth-

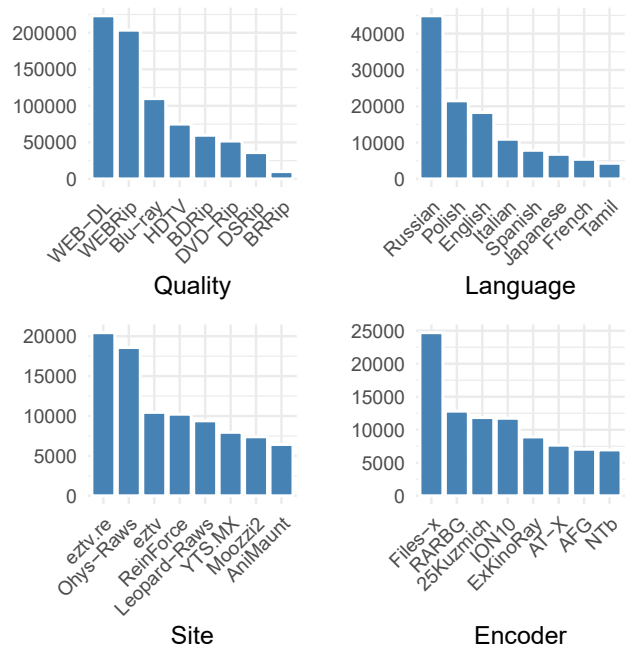

Figure 5: Select breakdowns of the matched video files: Quality labels, language tags, distribution "sites," and encoders (creators of the torrents themselves).

ers remain absent or only appear years after their release, we elaborate on this in our *Discussion*. This gives insight on the torrent network that other datasets and studies do not capture. We hypothesized that *MagnetDB* would develop large coverage of IMDb, but the data suggests that there are so many more small works that are not captured by our long-standing torrent crawler. It shows that there is so much more to the story on what gets torrented and what gets left out. This dataset challenges preconceived notions about the proliferation of pirated content and offers insights into what content is targeted by torrent suppliers.

Distribution

MagnetDB is FAIR (Wilkinson et al. 2016):

> OSF URL: https://osf.io/9eh47/
> DOI: https://doi.org/10.17605/OSF.IO/9EH47
>
> Academic Torrents:
> https://academictorrents.com/details/
> 1ce8202af7a500469177ed99de5cd9bf66078de0

Findable. *MagnetDB* maintains a persistent, globally unique identifier via DOI. The main page on Open Science Framework (OSF) describes the dataset in detail with rich metadata. Furthermore, our data can be shared on other platforms with attribution given its CC BY 4.0 license.

Accessible. We utilize an OSF data repository for *MagnetDB* and its successive versions. The protocol and data therein is open and free, and the data is permanently available for public use. For unauthenticated public release,

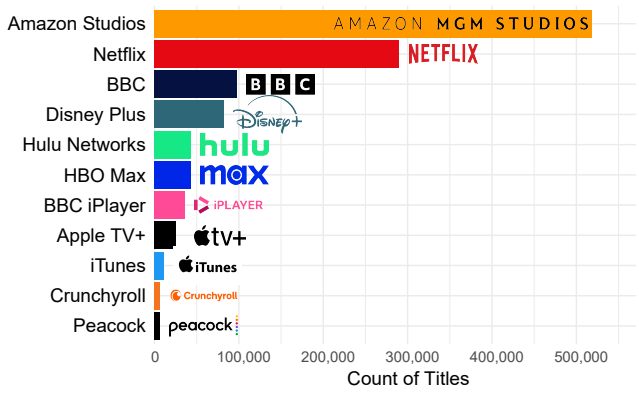

Figure 6: Distribution of torrent-matched titles by streaming service or production network. This data suggests that exclusive or region-locked content fosters higher piracy rates. All logos shown are trademarks of their respective owners.

we do not include the magnet links themselves, as *MagnetDB* contains content identifiers to a significant amount of pirated, copyrighted content. We do not wish to make it easier for a nefarious actor to leverage *MagnetDB* to find pirated content. However, for academic research purposes we will make the magnet links themselves available upon request to the corresponding author with an authenticated, university affiliated account on OSF.

Interoperable. The database is delivered as a SQLite database, a stable, cross platform, and one of the most widely deployed database architectures in the world (Gaffney et al. 2022). Additionally, we provide a **.csv** version of a smaller, curated subset of *MagnetDB* that focuses specifically on movies and TV shows that we can reasonably match to the IMDb database. This allows *MagnetDB* to be cross-referenced to other data sources, by torrent, by file, and by other content identifiers such as movie title, for example.

Re-usable. *MagnetDB* conforms with our domain-relevant community standards for research on torrents. We support re-usability of the dataset with our choice of license (CC BY 4.0), clear data provenance, and a breadth and depth of data attributes. We have designed and curated this dataset with the purpose of it being used as foundational to future work in this field. Notably, **updates will be made annually** to this dataset, with subsequent additions published on the OSF repository.

Discussion

Our introduction of *MagnetDB* demonstrates the feasibility and value of a multi-year, large-scale dataset that explicitly focuses on the supply side of BitTorrent. By indexing torrent metadata from a vast swath of the BitTorrent DHT, we have assembled a granular view of what is being uploaded and shared, rather than who downloads or just how swarms evolve. In doing so, our dataset illuminates broader social, cultural, and technical dynamics that underlie piracy ecosystems. Below, we discuss the key insights gleaned from *MagnetDB*, reflect on its ethical implications and limitations, and

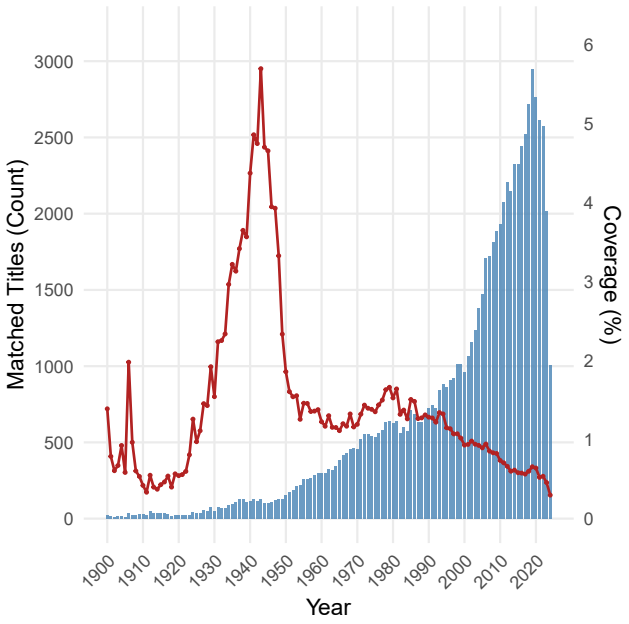

Figure 7: Coverage of *MagnetDB*'s matched subset by release year against IMDb. Blue bars correspond to the absolute count of matched titles per release year; the red line shows percentage coverage relative to IMDb.

consider pathways for future research using this resource.

Key Insight and Novel Contributions

Longitudinal and Content-Centric View. Where many existing studies concentrate on peer activity, swarm dynamics, or short-term snapshots of piracy, *MagnetDB* offers longitudinal perspective spanning over five years. This temporal breadth enables analyses of how supply-side behaviors evolve in parallel with technological, social, and even geopolitical factors. Our integration of file-level metadata with IMDb identifiers gives a richer picture of what exactly is getting uploaded and by whom. Crucially, these linkages enable targeted inquiries into the popularity and availability of specific media types, revealing distribution trends that cannot be captured by focusing solely on tracker-level measurements.

Supply-Side Cultural Dynamics. By zeroing in on the individuals and groups who create torrents, *MagnetDB* underscores the influential role of The Scene in shaping what content is made available and how. Our findings support the ida that The Scene's practices, however informal and federated, retain a high degree of consistency and organization. Encoders still label and structure torrents according to well-established naming conventions, a tradition upheld by a subcultural gift economy and its emphasis on reputation, speed, and quality. While The Scene's norms can appear obscure, they significantly impact the broader network. Groups with strong reputations set naming standards that other encoders replicate, demonstrating subcultural capital translates into soft power within this ecosystem. This is a level of analysis only provided by a dataset like *MagnetDB*.

Evidence of Selective Coverage. Although *MagnetDB* spans tens of millions of torrents, our analysis of matched video files reveals that a substantial fraction of the IMDb catalog is simply never torrented. We suspect this is particularly pronounced for obscure titles with low demand or limited mainstream attention based on what is known about how encoders and sites prioritize the availability of content. Despite ongoing expansions in film and television production, The Scene and associated groups prioritize specific, high-value content, often new releases or classic titles with enduring popularity. Our initial exploration of *MagnetDB* shows a diminishing proportion of new releases that appear in torrent form over time, which suggests a dynamic interplay between supply-side capacity, user demand, and underlying motivations in this gift economy. This raises intriguing questions about the cultural gatekeeping role performed by these piracy groups, who effectively determine which titles, and in which form, achieve broad accessibility through these P2P networks–exploration of these questions we leave for future work.

Limitations

Potential Bias in Coverage. The BitTorrent DHT is a constantly changing network, and any given vantage point can miss torrents that either have extremely short lifespans or exist in small, localized subgraphs of the network with low propagation to the rest of the network. While we took a longitudinal and consistent approach, no data collection pipeline can fully index every torrent in existence due to the decentralized topology of the P2P network. Our coverage is therefore best viewed as a robust, but not provably exhaustive, snapshot of real-world activity at scale.

Matching Imperfections. Torrent file names can, and often are, incomplete, misleading, or unhelpfully generic, which complicates any automated matching to external databases such as IMDb. Despite internal validation and conservative BM25 thresholds, spurious matches and missed titles inevitably persist. Additionally, due to how the BM25 algorithm calculates match scores, where the max possible score is positively correlated to the length of text, there is a potential for match bias towards entries whose titles are longer. Researchers using *MagnetDB* should remain aware of the potential for false positives or negatives and choose their own matching thresholds and conduct data validation checks appropriate for their specific use cases.

Ethical and Legal Implications. While it is well-known that BitTorrent hosts a wide array of copyrighted content, the platform is fundamentally a dual-use technology. It serves many legitimate purposes, such as disseminating open-source software or bypassing censorship in oppressive regimes. Our role in providing *MagnetDB* is to offer a dataset, that will become an updating catalog of datasets, that can advance empirical research on sociotechnical systems, not to endorse or facilitate copyright infringement or other crimes. We do not distribute magnet links themselves in the publicly accessible version of *MagnetDB*, balancing transparency for academic inquiry with a commitment to minimize unethical or illicit use.

Applications and Future Directions

Cultural Analytics. The structured metadata in *MagnetDB* supports diverse lines of inquiry into sociocultural trends. For instance, researchers could examine how films, television shows, and other media travel across linguistic and cultural borders over time. A potential avenue of inquiry involves analyzing which titles are dubbed, subtitled, tagged, or otherwise adapted for different languages. By tracking the emergence of multi-language torrents and the frequency of specific language tags or subtitles, researchers could investigate how certain genres, franchises, or classic works proliferate beyond their original audiences. This form of cultural diffusion reflects both grassroots demand (e.g., fans creating or sharing subtitles) and broader, industry-led strategies to reach new markets.

MagnetDB also facilitates longitudinal studies on the ebb and flow of cultural relevance. Some titles appear for a fleeting moment only to vanish as interest wanes. Others persist in active torrents long after their initial release, sustained by perennial fan enthusiasm, new adaptations, or nostalgic revivals. The dataset allows for comparative analyses of how, for instance, cult classics gradually gain renewed popularity or how once-obscure films become "resurrected" due to external triggers (e.g., anniversaries, streaming re-releases, or cultural rediscovery). By correlating torrent availability with societal events and production milestones, scholars can trace the lifecycle of media artifacts, identifying when and why certain works achieve evergreen status while others fade away or abruptly re-emerge.

Policy and Anti-Piracy Strategies. Given its multi-year breadth, *MagnetDB* can inform the design and evaluation of anti-piracy interventions. Industry groups and policymakers could, for example, assess whether certain enforcement actions or takedown efforts correlate with measurable changes in torrent supply, even from specific individuals, sites, or collectives. Likewise, our data could support modeling how new offering types, such as simultaneous digital releases, shift the supply of pirated content. In contrast to many existing data sources that rely on user-facing distribution sites, *MagnetDB* enables direct observation of the supply side.

Evolving Supply of Other Media. Beyond movies and TV shows, the dataset encompasses torrents for a wide variety of media, from e-books and music to software packages. These alternate modalities follow distinct production, distribution, and consumption practices. Software torrents, for instance, raise unique security concerns due to the risk of embedded malware, a fundamentally different threat model from video files. Studies could therefore examine how software piracy evolves over time, comparing it with other forms of digital media in terms of both scale and supply patterns. Similarly, examining e-book and audio content could yield fresh perspectives on cultural diffusion, user demand for niche or academic material, and shifting norms around digital rights management.

Additionally, *MagnetDB* captures torrents for a considerable volume of adult content, an area of increasing policy scrutiny in certain jurisdictions. Researchers and policymakers alike could leverage torrent-level evidence to explore how legislative or platform-specific bans on adult content affect its distribution in P2P networks. For instance, if newly passed age-verification laws or website blocks reduce access to certain sites, *MagnetDB* data may reveal whether BitTorrent sees a compensatory spike in adult content sharing. By mapping these shifts, scholars could assess the efficacy and potential unintended consequences of bans or restrictions that attempt to control online adult content.

Beyond BitTorrent: Emerging P2P Networks. Although BitTorrent remains one of the most widely used P2P protocols, newer networks, such as IPFS, are increasingly being used to distribute illicit content (Sokoto et al. 2024; Shi et al. 2024). Future work should leverage *MagnetDB* to cross-reference if these emerging platforms carry the same supply-side incentives and subcultural norms. The introduction of cryptocurrency-based incentives in file-sharing systems revitalizes an old avenue of research, as the traditional gift economy model may once again hybridize to give way to more profit-oriented operations.

Conclusion

By making *MagnetDB* openly available for academic use, we seek to catalyze a new wave of research on piracy, subcultural economies, and P2P data exchange. Our holistic approach, spanning torrent discovery, file-level parsing, and IMDb matching, provides a more comprehensive lens on what is being uploaded and who is uploading it. While digital piracy continues to prompt legal and ethical debates, it remains an enduring sociotechnical phenomenon with implications for information access, creative industries, and digital culture. We believe *MagnetDB* will prove a valuable springboard for those seeking to understand and interpret these complex dynamics through the lens of real-world, empirical data.

References

1337x. 2007. 1337x. https://1337x.to. Accessed: 2025-04-02.

Alper, B. M. 2020. Magnetico. https://github.com/boramalper/magnetico. Accessed: 2025-01-13.

Ancell, N. 2024. Torrenting copyrighted content is illegal, yet nearly half of US adults do it. https://cybernews.com/privacy/american-adults-torrent-movies-tv-shows. Accessed: 2025-01-12.

Basamanowicz, J.; and Bouchard, M. 2011. Overcoming the Warez paradox: online piracy groups and situational crime prevention. *Policy & Internet*, 3(2): 1–25.

Bindlish, D. 2016. Parse Torrent Name - Python Package. https://pypi.org/project/parse-torrent-name. Accessed: 2024-12-05.

BitTorrent. 2025. BitTorrent. https://www.bittorrent.com. [Accessed 10-January-2025].

Bode, K. 2018. The Rise of Netflix Competitors Has Pushed Consumers Back Toward Piracy. https://www.vice.com/en/article/bittorrent-usage-increases-netflix-streaming-sites. Accessed: 2025-01-12.

BTDigg. 2025. BTDigg: The BitTorrent DHT Search Engine. https://btdig.com. Accessed: 2025-01-12.

Danaher, B.; Smith, M. D.; and Telang, R. 2020. Piracy landscape study: Analysis of existing and emerging research relevant to intellectual property rights (IPR) enforcement of commercial-scale piracy.

David, M. 2017. *Sharing: crime against capitalism*. John Wiley & Sons.

Dementiev, E.; and Fenton, N. 2016. Bayesian Torrent Classification by File Name and Size Only. In *Conference on Probabilistic Graphical Models*, 136–146. PMLR.

Digital Music News. 2024. Verizon Faces Copyright Lawsuit From Major Labels. https://www.digitalmusicnews.com/2024/07/15/verizon-copyright-lawsuit-major-labels-july-2024. Accessed: 2025-01-12.

Ellis, G. 2023. Streaming has a consumer and a piracy problem; the answer lies in the music industry. Accessed: 2025-01-12.

Gaffney, K. P.; Prammer, M.; Brasfield, L.; Hipp, D. R.; Kennedy, D.; and Patel, J. M. 2022. Sqlite: past, present, and future. *Proceedings of the VLDB Endowment*, 15(12).

Goldman, E. 2005. The challenges of regulating Warez trading. *Social Science Computer Review*, 23(1): 24–28.

Hétu, D. D.; Morselli, C.; and Leman-Langlois, S. 2012. Welcome to the scene: A study of social organization and recognition among warez hackers. *Journal of Research in Crime and Delinquency*, 49(3): 359–382.

Huizing, A.; and van der Wal, J. A. 2014. Explaining the rise and fall of the Warez MP3 scene: An empirical account from the inside. *First Monday*.

IMDb. 2024. IMDb Non-Commercial Datasets. https://developer.imdb.com/non-commercial-datasets. Accessed: 2024-12-05.

Internet Archive. 2023. Torrent Metadata Archive Sample. https://archive.org/details/torrent_metadata_archive_sample. Accessed: 2025-01-12.

Kamphuis, C.; De Vries, A. P.; Boytsov, L.; and Lin, J. 2020. Which BM25 do you mean? A large-scale reproducibility study of scoring variants. In *42nd European Conference on IR Research (ECIR 2020)*, 28–34. Springer.

Kiwi Torrent Research. 2023a. Kiwi Torrent Research - Uncensored BitTorrent Data Set. Accessed: 2025-01-12.

Kiwi Torrent Research. 2023b. Kiwi Torrent Research: Exploring Peer-to-Peer Network Dynamics. https://github.com/Kiwi-Torrent-Research/Kiwi-Torrent-Research. Accessed: 2025-01-12.

Le Blond, S.; Legout, A.; Le Fessant, 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 *3rd USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET'10)*.

Lifewire. 2024a. Gen Z and Millennials Are Leading a New Wave of Streaming Piracy—Here's Why. https://www.lifewire.com/content-privacy-survey-8739047. Accessed: 2025-01-12.

Lifewire. 2024b. Streaming Prices Are Out of Control and It's Pushing People Back to Pirating. https://www.lifewire.com/high-streaming-prices-bring-back-pirating-8717968. Accessed: 2025-01-12.

LimeTorrents. 2009. LimeTorrents. https://www.limetorrents.info. Accessed: 2025-01-12.

McCandless, D. 1997. Warez wars. https://www.wired.com/1997/04/ff-warez. Accessed: 2025-01-12.

Pouwelse, J.; Garbacki, P.; Epema, D.; and Sips, H. 2005. The bittorrent p2p file-sharing system: Measurements and analysis. In *International Conference on Peer-to-Peer Systems (IPTPS)*, 205–216. Springer.

RARBG. 2008. RARBG. https://rarbg.to. Accessed: 2025-01-12.

Rehn, A. 2004. The politics of contraband: The honor economies of the warez scene. *The journal of socio-economics*, 33(3): 359–374.

Robertson, S.; Zaragoza, H.; et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4): 333–389.

Shi, R.; Cheng, R.; Han, B.; Cheng, Y.; and Chen, S. 2024. A Closer Look into IPFS: Accessibility, Content, and Performance. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 8(2): 1–31.

Sokoto, S.; Balduf, L.; Trautwein, D.; Wei, Y.; Tyson, G.; Castro, I.; Ascigil, O.; Pavlou, G.; Korczyński, M.; Scheuermann, B.; et al. 2024. Guardians of the Galaxy: Content Moderation in the InterPlanetary File System. In *USENIX Security Symposium*, 1507–1524.

Su, M.; Zhang, H.; Du, X.; Fang, B.; and Guizani, M. 2013. A measurement study on the topologies of BitTorrent networks. *IEEE Journal on Selected Areas in Communications*, 31(9): 338–347.

The Pirate Bay. 2003. The Pirate Bay. https://thepiratebay.org. Accessed: 2025-04-02.

US Chamber of Commerce. 2019. Impacts of Digital Piracy on the US Economy. https://www.uschamber.com/technology/data-privacy/impacts-of-digital-piracy-on-the-u-s-economy. Accessed: 2025-01-12.

Wang, H.; Liu, J.; and Xu, K. 2010. Measurement and enhancement of BitTorrent-based video file swarming. *Peer-to-peer Networking and Applications*, 3: 237–253.

Wang, L.; and Kangasharju, J. 2013. Measuring large-scale distributed systems: case of bittorrent mainline dht. In *IEEE P2P 2013 Proceedings*, 1–10. IEEE.

WebTorrent. 2025. WebTorrent: Streaming Torrent Client. https://webtorrent.io. Accessed: 2025-01-12.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9.

Wolchok, S.; and Halderman, J. A. 2010. Crawling BitTorrent DHTs for Fun and Profit. In *4th USENIX Workshop on Offensive Technologies (WOOT 10)*.

Zhang, C.; Dhungel, P.; Wu, D.; and Ross, K. W. 2011. Unraveling the BitTorrent ecosystem. *IEEE Transactions on Parallel and Distributed Systems*, 22(7): 1164–1177.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. Our database is aggregated from publicly available magnet links and torrent metadata; no personally identifiable information is collected or released. Potentially sensitive elements (e.g., direct links) are withheld from public release to reduce misuse.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes. We clearly state that we introduce a large-scale torrent dataset and document its composition, methodology, and potential uses.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. We explain our approach for collecting torrents, parsing file metadata, and matching titles to IMDb.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. We discuss naming conventions, potential biases in content availability, and how "the scene" culture might affect representativeness.
   (e) Did you describe the limitations of your work? Yes. We have a "Limitations" section that addresses potential biases, incomplete coverage, and risk of partial false positives in matching.
   (f) Did you discuss any potential negative societal impacts of your work? Yes. We acknowledge that torrent data is often associated with pirated content and discuss ethical considerations.
   (g) Did you discuss any potential misuse of your work? Yes. We note that while we do not release magnet links publicly, we remain alert to risks of facilitating unauthorized content access.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. We describe our access-control policy for magnet links, omit direct links in the public release, and provide a documented process for responsible access.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes. We cite IMDb for the metadata, Magnetico for torrent indexing, and other relevant software libraries.
   (b) Did you mention the license of the assets? Yes. We clarify IMDb data usage terms and open-source licenses for any libraries.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes. We provide data documentation plus a link to the OSF repository and Academic Torrents link that has all the data.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? We use publicly available torrent metadata and do not include PII, so explicit consent is not applicable.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. We explicitly note there is no PII and disclaim that while some files may have offensive or adult material, the content itself is not distributed.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes. *Distribution* section details how we follow FAIR principles.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes. In addition to what is covered in the main body of the paper, we include full dataset documentation in the `README` on the data repository.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? No IRB required, as we do not collect data directly from human subjects.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Appendix

Examples of a Low and High Match Score

Low Match Score

Filename: `Riviera.S02E01.WEBRip.x264-ION10.mp4`
Candidate Match: *On the Riviera*
Correct Title: *Riviera*
Match Score: 18.3
A recurring problem arises from short prepositions and articles (e.g., "on," "the") that mismatch the true title.

High Match Score

Filename: `shadowhunters.the.mortal.instruments.s03e20.1080p.web.h264-tbs.mkv`
Matched Title: *Shadowhunters: The Mortal Instruments*
Match Score: 260
This example highlights a near-ideal match, where the full series title appears in the torrent name, leading to a high BM25 score.