Cross-Artifact Comparative Analysis of Legal and Illegal Korean Streaming Sites

Su-bin Lee[1], Da-hyung Kim[1], Jeong-hwa Ryu[1], and Seongmin Kim.[1]

Department of Convergence Security Engineering, Sungshin Women's University
{20222605, 20231065, 20231093, sm.kim}@sungshin.ac.kr

Abstract

Illegal streaming platforms have proliferated alongside the rapid expansion of online video streaming services. These sites provide unauthorized access to copyrighted content while mimicking the appearance of legitimate Over-the-Top (OTT) platforms, thereby undermining legal services and exposing users to various risks. In this study, we focus on Korean streaming platforms, conducting a cross-artifact comparative analysis of real-world legitimate and illegal sites across three categories of Web artifacts: HTML structure, cache entries, and cookies. Our analysis shows that legitimate platforms maintain consistent domain-based resource calls and employ standardized cookie management practices. In contrast, illegal sites exhibit distinctive and immature operational patterns, including extensive use of external domains, repeated reliance on default session management cookies, opaque affiliations with advertising networks, and high redundancy in cache data. Moreover, some pairs of illegal sites displayed strong similarities across all three artifact types, suggesting the possibility of shared infrastructure or common operators. By revealing these consistent cross-artifact patterns, this study provides empirical evidence that can inform identification of illegal streaming platforms.

Keywords— Illegal Streaming Sites, Digital Forensics, Cross-artifact, and Web artifacts.

1 Introduction

With the rise in internet usage and the proliferation of various online video streaming services, competition among Over-the-Top (OTT) platforms has intensified. Major services like Netflix, TVING, and Coupang Play, continuously invest in technology to deliver high-quality content and enhance user experience. However, the spread of illegal streaming sites has emerged alongside this rapid growth. These sites provide copyrighted material without authorization, bypass digital rights management (DRM), and erode the business models of legitimate providers. According to the Copyright Act (Article 124) and the Supreme Court's full panel ruling (2017Do19025) [1] in South Korea, operating such platforms constitutes copyright infringement. Moreover, systematic linking practices are regarded as aiding and abetting unauthorized public transmission.

Despite the growing prevalence of illegal streaming platforms, existing studies have primarily concentrated on surface-level indicators such as domain names or network traffic patterns [2, 3]. While these approaches provide useful insights into access routes and traffic behaviors, they fall short of capturing the internal composition of sites. In particular, HTML structures have often been treated merely as supplementary evidence, and little systematic attention has been given to semi-static artifacts, such as caches or cookies that record user interaction traces. As a result, the broader operational characteristics of illegal platforms remain insufficiently understood.

This study addresses this gap by conducting a cross-artifact comparative analysis between legal and illegal streaming sites. Specifically, we analyze structural and operational differences across three dimensions: 1) HTML structure, 2) Web cache entries, and 3) cookie usage patterns. Through this multi-faceted approach, we reveal distinctive characteristics of illegal sites, such as reliance on low-cost and weakly regulated domains, repeated dependence on default session management cookies, and extensive embedding of external advertising or tracking resources. In contrast, legal platforms demonstrate standardized, transparent, and internally managed practices across all artifacts examined.

In summary, our contributions are as follows:

- We present a comprehensive cross-artifact comparison that extends beyond HTML to include cache and cookie data.
- We demonstrate that illegal streaming platforms rely on immature and ad hoc operational practices, which sharply contrast with the standardized approaches of legitimate OTT services.
- We show that these cross-artifact patterns can serve as practical criteria for identification and forensic investigation of illegal platforms, with implications for enhancing Web security and copyright protection.

2 Related Work and Motivation

Existing research on the ecosystem of illegal streaming sites has primarily focused on domain-level identification or network traffic analysis [2, 3]. Although such approaches are useful for examining access paths or traffic characteristics, they remain limited because they overlook internal page structures, user interaction aspects, and broader Web artifacts. Some studies[4] have analyzed various metrics within the Illegal Movie Streaming Services (IMSS) ecosystem, including DOM structure and cookies, but they did not perform analysis on the media player original source—a factor that could indicate the same operator.

Moreover, while some studies [5] have directly considered HTML structures, most treat them only as auxiliary indicators (e.g., analyzing URL patterns or redirection paths) rather than as central analytical elements. As a result, comprehensive investigations into conventional Web artifacts (e.g., cache and cookie) have received relatively little attention.

Accordingly, we aim to fill this gap by systematically comparing structural differences between legal and illegal streaming sites. Unlike prior work, our analysis extends beyond HTML to include diverse Web artifacts, such as caches and cookies. In particular, we provide deeper insight into the internal data-processing mechanisms of illegal sites, highlighting distinctive structural patterns, differences from legal OTT services, and even indications of sites hosted by the same provider. In doing so, we establish a foundation for developing identification techniques and practical security response strategies against illegal streaming platforms.

For our analysis, we selected four legal streaming sites and fifteen illegal streaming sites. To collect the HTML, we accessed each service with a designated piece of content and recorded the HTML while playing that content. Note, similar content was chosen instead if identical content was unavailable. Also, cache entries and cookies generated on local PCs during visits to both legal and illegal streaming sites were also collected. The complete list of streaming sites analyzed in this study is provided in Table 1.

Table 1: List of target streaming sites

Category	Site Names
OTT sites (Legal)	Netflix, TVING, Wavve, Coupang Play
Illegal sites	noonoo TV, TVRoom, TVMoaa, WangTV, TVHOT, TVChak, TVWIKI, Aniweek, ANIWOLF, BozayoNet HOOHOO TV, aniLIFE, Byulbyul TV, TVMON, Linkkf

3 Comparative Analysis of the HTML Structure

In this section, we present the results of our static HTML analysis. The collected HTML documents were categorized by tag and then broadly classified according to six criteria, with the analysis emphasizing structural differences across legal and illegal streaming sites: 1) evasion domain structuring, 2) sharing and reuse of identical player links, 3) mimicking of legal streaming sites, 4) number of external embeddings, 5) commonly accessed API domains, and 6) advertising banner insertion patterns and parameters.

3.1 Evasion Domain Structuring

Illegal streaming sites frequently change their domain addresses, often using easily replaceable and less-regulated domains (e.g., .tv, and .xyz) to evade enforcement actions. This strategy allows operators to quickly relaunch sites under new addresses shortly after takedowns. In addition, these sites commonly provide alternative communication channels, such as Telegram (t.me) links or informational pop-ups, to redirect users if the original site is blocked, ensuring continuous accessibility and user retention. Overall, illegal streaming platforms employ a variety of tactics to create workarounds and alternative access methods, enabling them to sustain user traffic and remain operational despite enforcement measures.

3.2 Sharing and Reuse of Identical Player Links

Most legal streaming sites use custom-built video players that rarely rely on iframe tags, loading resources directly from their own domains and linking JavaScript paths internally. In contrast, illegal streaming sites frequently embed external video players using iframe and script tags or reuse player resources across multiple domains. Specifically, 11 out of 15 illegal streaming sites analyzed retrieved videos externally via iframe, with some sharing identical player domains or HTML structures, indicating potential code-sharing or common management. For example, identical HTML structures appear within iframe tags across the same domain (Figure 1) and external domains match precisely down to the video URL paths (Figure 2) across illegal sites. These findings suggest centralized distribution of illegal streaming content.

```
#TVHOT, TVWIKI
<div class="embed-container">
   <iframe id="view_iframe" width="560" height="315" scrolling="no"
    src="https://player.bunny-frame.online/?s=4&amp;..."
   frameborder="0" allow="accelerometer; autoplay; clipboard-write; encrypted-media;
   gyroscope; picture-in-picture; web-share" allowfullscreen="">
   </iframe>
</div>
```

Figure 1: Suspected reused code in different illegal sites

3.3 Mimicking of Legal Streaming Sites

Some illegal streaming sites directly reference content assets from legal services or closely mimic their structures to build user trust and obscure their illicit nature. In addition, these sites frequently load external style sheets and scripts from services (e.g., fonts.googleapis.com and cloudflare.com) to closely replicate the UI layouts of legal platforms. Notably, some illegal sites reproduce not only unauthorized content but also the full HTML structure and JavaScript call paths of legal services, representing a deeper form of infringement.

3.4 Number of External Embeddings

Typically, legal streaming sites have limited external JavaScript or iframe calls, with scripts for statistical analysis or user authentication loaded primarily from trusted domains (e.g., googletagmanager.com and instagram.com). For example, Wavve uses only four external domains, most of which are

```
#Aniweek
<div id="movie_box">
     <div id="movie_player" class="embed-responsive embed-responsive-16by9 mcpalyer">
         <iframe src="https://michealcdn.com/video/60cab8d...c0b872b"
         width="640px" height="360px" frameborder="0" allowfullscreen=""></iframe>
    </div>
</div>

#ANIWOLF
<div class="row detail-player" style="display:none;">
     <div class="col p-0">
          <div class="ratio ratio-16x9">
                 <iframe src="https://michealcdn.com/video/60cab8d...c0b872b"
                 allowfullscreen=""></iframe>
             </div>
     </div>
</div>
```

Figure 2: Same external player domain address in different illegal sites

confirmed to be scripts for statistical analysis or login authentication with clear purposes and origins. In contrast, illegal sites freely insert numerous external domains, frequently including ambiguous or advertising-oriented domains. Specifically, domains like Byulbyul TV, noonoo TV, and TVWIKI irregularly called images, ads, iframes, and tracking codes through over 20 external domains. Such structures may facilitate extensive user tracking, enable aggressive ad revenue generation, or even provide channels for malicious code injection. Table 2 summarizes the result.

Table 2: External domain usage of legal and illegal streaming sites

Category	Site Names	# of external domains	External domain examples
OTT sites (Legal)	Coupang Play	8	instagram.com, schema.org, etc.
	TVING	6	googletagmanager.com, ad.doubleclick.net, etc.
	Wavve	4	googletagmanager.com, ad.doubleclick.net, etc.
	Netflix	1	github.com
Illegal Sites	Byulbyul TV	55	bet38join.com, winner1.site, 1bet1.one, 1000-new.com, pan-8282.com, etc.
	noonoo TV	45	rb-002.com, son-509.com, 1bet1.sc, bsbs-777.com, rd-365.com, etc.
	TVWIKI	25	ppt-002.com, cms-002.com, rb-002.com, t.me, toss.im, bet38join.com, etc.

3.5 Commonly Accessed API Domains

Legal streaming sites primarily utilize their own domains for API calls, with minimal external resource access. For instance, TVING retrieves images and data through internal API endpoints such as image.tving.com and stillshot.tving.com, rarely engaging external domains. In general, legal streaming platforms maintain a simple, stable domain structure, mostly interacting only with official, trusted domains, such as instagram.com and twitter.com, as well as advertising and analytics services (e.g., googletagmanager.com).

In contrast, illegal streaming sites exhibit broadly distributed external API call patterns. Notably, the domain t.me (Telegram) appeared in 8 of the 15 analyzed sites, presumably to redirect users in case of site blocking or domain changes. In addition, numerous sites load thumbnail images and video resources from obscure external sources (e.g.,img-requset4.digitalori3nx.com and images2.imgbox.com). Moreover, some sites embed scripts from analytics and user-tracking domains, such as histats.com and shinystat.com, likely to optimize advertising revenue through traffic management and behavioral analysis. Despite seemingly different operational entities, these illegal sites share common external domain structures, suggesting resource sharing or centralized management.

3.6 Advertising Banner Insertion Patterns and Parameters

An in-depth examination of the four legal streaming services shows that these platforms do not rely on external advertising domains operated by third parties. Even when advertisements are displayed, they are limited to official partners or the platforms' own domains. It is worth noting that no evidence of external redirection practices was observed, such as numeric–alphabetic mixed domains that appear auto-generated or URL paths like `/redirect/`.

In contrast, illegal streaming sites intersperse presumed advertising links throughout their HTML code, often embedded between primary navigation paths (e.g., `/movie` and `/drama`). In several cases, these sites redirect traffic to external domains through `/redirect/` paths. The links frequently point to gambling platforms, adult content, or Telegram channels, and they are typically structured with referral parameters, such as `?code=`, `?ref=`, and `?regcode=`. Notably, the domains themselves often combine numbers and letters (e.g., ppt-002.com and rb-002.com), strongly suggesting that these URLs are automatically generated by advertising networks [6]. These contrasting patterns highlight the fundamental differences in advertising practices between legal and illegal streaming sites, underscoring how the latter rely on aggressive and deceptive techniques that can serve as reliable forensic indicators.

4 Cache Analysis

We also examined the structural characteristics of cache entries to compare legal and illegal streaming sites, with particular attention to their reliance on external resource calls. The collected cache data were analyzed in terms of URL frequency vectors, enabling us to identify structural similarities across sites and detect distinct dependency patterns. A Web cache temporarily stores website resources in local storage during browsing sessions to improve page loading performance. From a forensic perspective, this characteristic makes cache data especially valuable, as it retains records of specific site visits and external resource calls, thereby offering useful clues for reconstructing user behavior.

Illegal streaming sites, in particular, exhibit heavy reliance on third-party resources, such as advertising banners, user tracking scripts, and external video players. Consequently, URLs recorded in cache entries can serve as important indicators of site structure and operational dependencies. On the other hand, legal streaming platforms primarily deliver key assets, such as thumbnails or video players, directly from their own domains. To ensure clarity in comparative analysis, such first-party resources were excluded, allowing us to focus exclusively on dependencies involving third-party domains. For similarity evaluation, we applied two complementary metrics: 1) *Cosine Similarity*, which measures the angle between URL frequency vectors, and 2) *Jensen–Shannon Distance*, which quantifies the divergence between their probability distributions.

In this study, following the standard procedure for statistical outlier detection [7], the mean and standard deviation of similarity among legitimate platforms were set as the baseline. Cosine similarity measures the directional alignment between two URL frequency vectors; values closer to 1 indicate higher similarity between the two vectors. Conversely, Jensen–Shannon Divergence (JSD) represents the difference between two probability distributions; values closer to 0 indicate similar distributions. Using the mean and standard deviation of similarity among legitimate platforms, the threshold values were calculated: 1.234 for Cosine Similarity and 1.1337 for JSD. This statistical validation assessed the overall difference between the two similarity calculation methods and determined whether one method was statistically significantly superior to the other. Furthermore, the Mann–Whitney U test[8] revealed a highly statistically significant difference between the two groups ($U = 10706.0$, $p < 0.001$), and the effect size calculated using Cohen's d was also very large at 4.47. Correlation analysis also revealed Pearson's correlation coefficient and Spearman's correlation coefficient of –0.934 ($p < 0.001$) and –0.974 ($p < 0.001$), respectively, confirming a very strong negative correlation between the two variables.

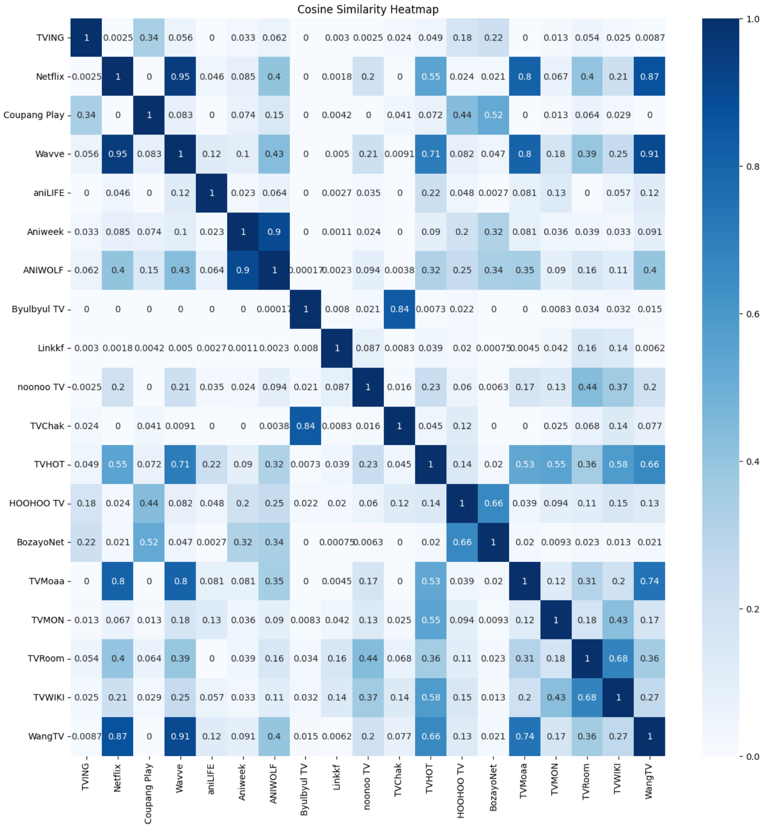


Figure 3: Cosine Similarity Heatmap

Cosine Similarity measures the angle between two vectors, where values closer to 1 indicate higher similarity. As shown in Figure 3, our analysis revealed notable similarities among several pairs of illegal streaming sites: Byulbyul TV–TVChak (0.84), TVMoaa–WangTV (0.74), and HOOHOO TV–BozayoNet (0.66). Particularly, ANIWOLF and Aniweek, previously identified as using the same external player in the HTML analysis, showed an exceptionally high similarity of 0.90. Likewise, TVHOT and TVWIKI, where iframe code reuse was suspected, exhibited a relatively high similarity score of 0.58. Conversely, WangTV displayed unexpectedly high similarity with multiple legal sites, ranging from 0.81 to 0.91. This anomaly is interpreted as a consequence of simplified external call patterns, likely because the site either excluded advertising banners altogether or handled video playback through separate windows rather than embedded structures.

These high cosine similarity scores can be primarily attributed to the repeated use of identical external player components, ad-serving scripts, and CDN endpoints. In essence, sites that reuse the same third-party infrastructures generate nearly parallel URL frequency vectors, leading to elevated cosine similarity. Conversely, outlier cases such as WangTV, which show high similarity with legal platforms, likely result from simplified or minimalistic cache patterns rather than genuine structural overlap.

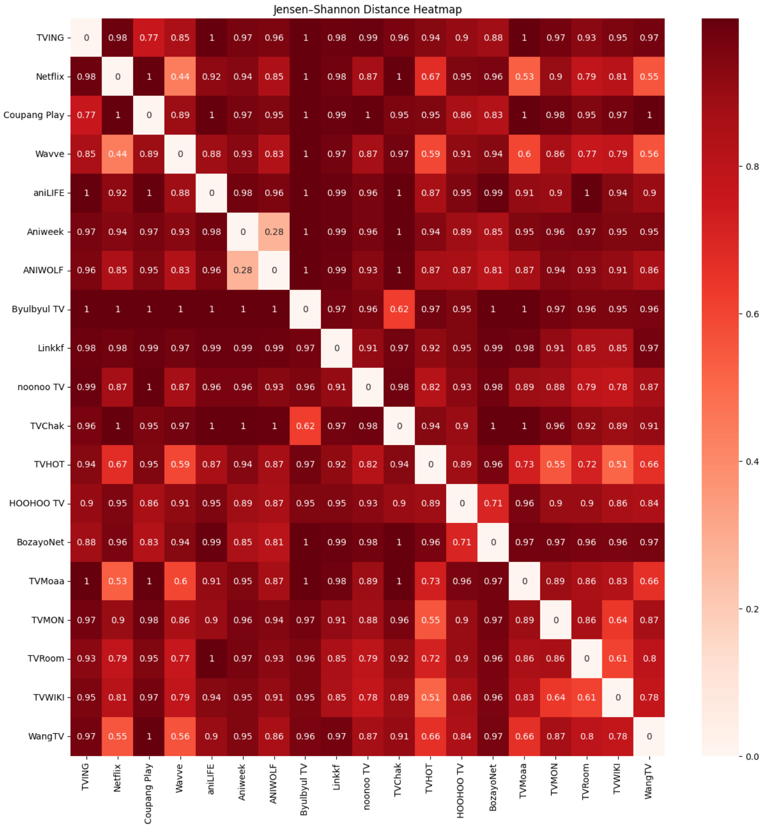


Figure 4: Jensen-Shannon Distance Heatmap

Jensen–Shannon Distance measures the divergence between two probability distributions, where values closer to 0 indicate greater similarity. The overall results are depicted in Figure 4. Our analysis showed moderate similarity across several pairs of illegal streaming sites: TVHOT–TVWIKI (0.51), TVHOT–TVMON (0.55), TVRoom–TVWIKI (0.61), Byulbyul TV–TVChak (0.62), and TVMoaa–WangTV (0.66). Again, ANIWOLF and Aniweek exhibited a particularly low distance of 0.28, suggesting a high degree of similarity between the two sites. As illustrated in Figure 5, both domains frequently called resources from the `mars-star-nd4.xyz` and `n4player.xyz` families, indicating shared reliance on the same advertising and streaming infrastructure.

Overall, both Cosine Similarity and Jensen–Shannon Distance consistently revealed structural overlaps among illegal streaming sites, with particularly high similarity observed in pairs that shared external video players or iframe structures, as previously identified in the HTML analysis. These findings suggest that illegal streaming platforms frequently reuse identical codebases or depend on the same third-party infrastructure, thereby reducing their operational diversity and increasing detectability. However, relying solely on such similarity metrics makes it challenging to establish precise classification criteria, as certain anomalies (e.g., WangTV's similarity to legal sites) highlight the potential for

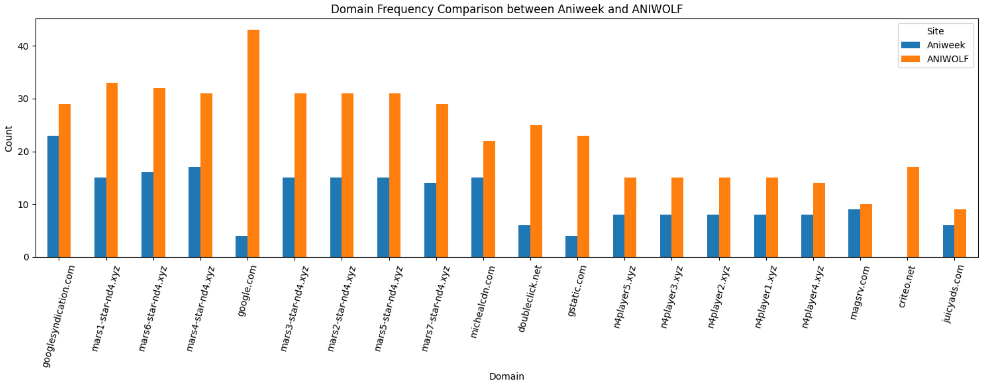


Figure 5: Distribution of URLs for “Aniweek” and “ANIWOLF”, limited to the top 20 domains.

misleading results. This underscores the necessity of incorporating additional Web artifacts, such as cookies or cross-site tracking identifiers, into the analysis.

5 Cookie Analysis

In addition, we analyzed differences between legal and illegal streaming sites with a focus on session cookies, advertising/tracking cookies, and third-party cookies. Cookies are small data files issued by Web servers and stored on users' devices for a limited duration; browsers automatically attach these cookies to subsequent requests to the originating server, enabling server-side recognition of returning users. This mechanism underpins user-specific state management and personalization, including shopping-cart persistence or login session maintenance. Additionally, authentication cookies are used for security purposes [9]. By examining the presence, frequency, and redundancy of particular cookies on illegal streaming sites, we aim to infer their operational practices and surface associated security and privacy vulnerabilities.

5.1 Session Management Cookies

Session cookies link a user's logged-in account to a specific session or ensure that the user is recognized consistently when navigating between pages. They are automatically deleted when the browsing session ends and are primarily associated with user authentication and short-term account access management. In contrast, persistent cookies remain stored in the user's browser until their pre-set expiration date and are typically used to retain long-term visit history or user preferences [10, 11]. Legal streaming sites typically use session management cookies (e.g., session_id, accessToken, authToken) to perform user authentication and session maintenance functions. This is an essential procedure in service provision, supporting legitimate functions, such as maintaining login status and providing personalized services per user.

However, illegal streaming sites often simplify this cookie management and lack sufficient security considerations. As shown in Table 3, the session management cookie `PHPSESSID` was commonly found on six out of eight investigated illegal streaming sites, whereas no evidence of its usage was observed in legal streaming platforms. According to the cookie definition database[12], `PHPSESSID` is a default session cookie used in PHP-based web applications. While the presence of `PHPSESSID` alone does not necessarily indicate a vulnerability, insecure implementations may expose sites to session fixation attacks [13]. The widespread presence of `PHPSESSID` among illegal sites therefore highlights their operational immaturity, illustrating a lack of investment in proper session handling practices. This structural difference serves as a clear marker distinguishing illegal platforms from legal services.

Table 3: Common Session Management and Authentication Cookie Usage by Illegal Streaming Sites

Cookie Name	Description	Sites
PHPSESSID	A cookie that manages user sessions; commonly used on PHP-based websites.	noonoo TV, TVMoaa, WangTV, TVHOT, HOOHOO TV, BozayoNet, Linkkf, Byulbyul TV, TVMON
cf_clearance	A Cloudflare security cookie used to maintain access to a site after passing Cloudflare's verification checks	noonoo TV, TVMoaa, BozayoNet, HOOHOO TV
c_ref_	A cookie that tracks user navigation history to manage sessions	noonoo TV, TVMoaa
AEC	A Google cookie used to detect fraud and secure account sessions	ANIWOLF, Aniweek
DV	A short-lived Google Ads cookie for session-based ad personalization	ANIWOLF, Aniweek
test_cookie	A DoubleClick cookie used to check browser cookie support	ANIWOLF, Aniweek

Similarly, the session management cookie cf_clearance was repeatedly observed on four of the eight investigated illegal streaming sites. The cf_clearance cookie is normally associated with Cloudflare's access verification process, where it is issued to confirm that a user is not a bot or to bypass specific access restrictions [14]. While session management cookies are employed by both legal and illegal platforms, the frequent use of cf_clearance in illegal streaming sites suggests a different operational pattern. Rather than functioning solely as a standard access verification mechanism, its repeated presence may indicate reliance on third-party challenge processes to regulate traffic or attempts to mask automated requests as legitimate user activity.

In addition to PHPSESSID and cf_clearance, other cookies such as c_ref_, AEC, DV, and test_cookie were also commonly observed among illegal streaming sites. The c_ref_ cookies function similarly to "call by reference" in programming, where the address of a variable is passed to a function, allowing direct modification of the original data. [15]

The AEC cookie is primarily used to detect spam, fraud, and abuse in advertising systems, ensuring that advertisers are not charged for invalid interactions and that content creators are fairly remunerated. [16]

DV cookies, session-based trackers used in advertising optimization, have a very short lifespan of one day and serve to monitor user interactions across platforms. Although temporary, their presence on illegal sites suggests reliance on third-party analytics or ad-related tracking mechanisms instead of internally managed session handling. [17]

Finally, the test_cookie is a marketing cookie used to verify whether a browser supports cookies, commonly deployed in Google Ads systems. While seemingly benign, its presence alongside other tracking cookies highlights the ad-centric or externally reliant operational patterns of illegal streaming sites. [18]

Collectively, these cookies—along with PHPSESSID and cf_clearance—illustrate simplified or externally dependent session and tracking management practices, underscoring the operational immaturity and insufficient security measures of illegal platforms. The configuration and presence of these cookies can therefore serve as reliable structural markers to distinguish illegal streaming services from legitimate ones.

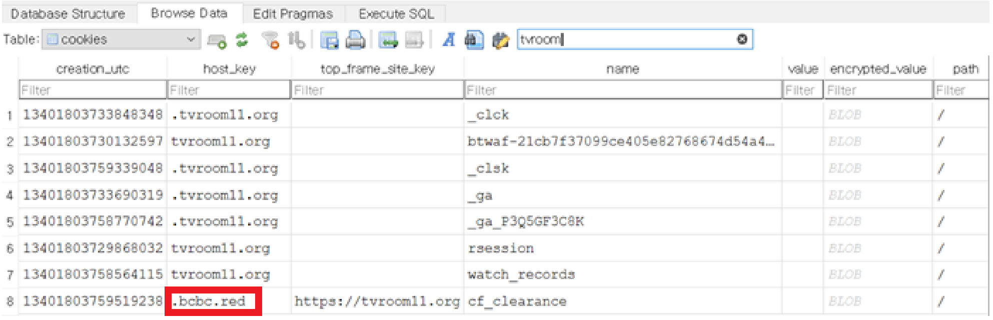
5.2 Advertising Tracking Cookies and Third-party Cookies

A clear distinction between legal and illegal streaming sites emerges in their use of third-party and advertising tracking cookies. Cookies can be broadly classified into first-party and third-party types depending on their issuing domain. First-party cookies originate from the visited site and support essential functions such as session management and login maintenance. Third-party cookies, by contrast, are issued by external domains through integrated services, such as advertising networks or analytics platforms, enabling targeted advertising and cross-site tracking of user behavior. When combined with tracking cookies, they can be used to record browsing activities and build behavioral profiles.

Our analysis shows that legal streaming platforms primarily set third-party cookies from reputable providers, including Google (_ga and __gads), DoubleClick, and Facebook (fbp). These are generally used for advertising performance monitoring and audience targeting. In contrast, illegal streaming sites frequently employed third-party cookies linked to low-credibility or obscure advertising networks. Such reliance not only facilitates aggressive tracking but also creates vectors for malicious ad injection, Potentially Unwanted Program (PUP) distribution, and unauthorized collection of session-related information.

In summary, legal sites typically rely on established networks for limited advertising and analytics purposes, whereas illegal sites employ cookies from unverified third-party domains. Therefore, the presence and origin of third-party cookies provide a clear differentiator between legal and illegal streaming sites, highlighting how immature and opaque cookie practices characterize illegal platforms and offer a reliable basis for their automated detection. We now present case studies of key third-party cookies identified in this study, focusing on their sources, the trustworthiness of their domains, and the security implications associated with their use.

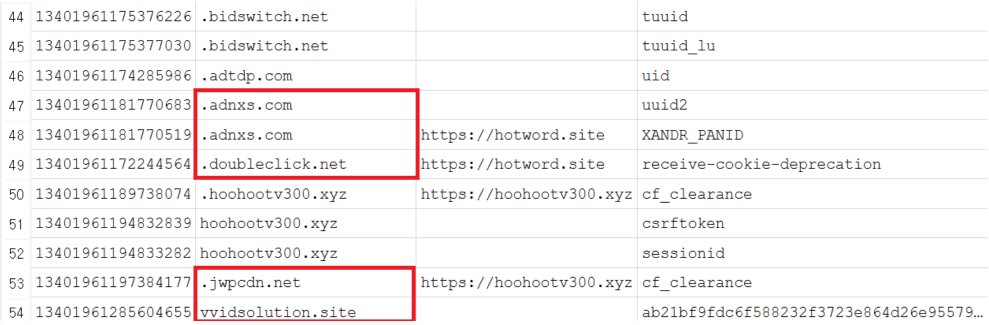
Case 1: `.bcbc.red` on "TVRoom". Traces of third-party cookies issued by the `.bcbc.red` domain were identified on the illegal streaming site "TVRoom.", as shown in Figure 6. Because there is almost no publicly available information about this domain, it appears to represent an advertising network of uncertain origin. Such reliance on opaque and unverified providers illustrates the immaturity of cookie management practices on illegal platforms. Unlike legal services, which typically employ established third-party networks, the use of obscure domains reflects a lack of operational transparency and standardization. .



	creation_utc	host_key	top_frame_site_key	name	value	encrypted_value	path
	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	13401803733848348	.tvrooml1.org		_clck		BLOB	/
2	13401803730132597	tvrooml1.org		btwaf-21cb7f37099ce405e82768674d54a4…		BLOB	/
3	13401803759339048	.tvrooml1.org		_clsk		BLOB	/
4	13401803733690319	.tvrooml1.org		_ga		BLOB	/
5	13401803758770742	.tvrooml1.org		_ga_P3Q5GF3C8K		BLOB	/
6	13401803729868032	tvrooml1.org		rsession		BLOB	/
7	13401803758564115	tvrooml1.org		watch_records		BLOB	/
8	13401803759519238	.bcbc.red	https://tvrooml1.org	cf_clearance		BLOB	/

Figure 6: Cookies from 'TVRoom'

Case 2: `.jwpcdn.net` and `.xyz` on "HOOHOO TV". Third-party cookies such as `jwpcdn.net` and `vvidsolution.site`, frequently observed on illegal streaming sites, originate from external domains that support video content delivery. On "HOOHOO TV" (in Figure 7), the cookie from `.jwpcdn.net` is presumed to be associated with JWPlayer's content delivery network (CDN), typically used to monitor playback performance or analyze user viewing patterns. While such services can enhance video delivery efficiency, their appearance on illegal platforms points to a reliance on external infrastructures rather than integrated, service-specific solutions. This dependency again reflects the immature operational practices of illegal streaming sites: instead of investing in proprietary or officially contracted CDN services, they often resort to easily accessible but weakly managed third-party resources [19].



44	13401961175376226	.bidswitch.net		tuuid
45	13401961175377030	.bidswitch.net		tuuid_lu
46	13401961174285986	.adtdp.com		uid
47	13401961181770683	.adnxs.com		uuid2
48	13401961181770519	.adnxs.com	https://hotword.site	XANDR_PANID
49	13401961172244564	.doubleclick.net	https://hotword.site	receive-cookie-deprecation
50	13401961189738074	.hoohootv300.xyz	https://hoohootv300.xyz	cf_clearance
51	13401961194832839	hoohootv300.xyz		csrftoken
52	13401961194833282	hoohootv300.xyz		sessionid
53	13401961197384177	.jwpcdn.net	https://hoohootv300.xyz	cf_clearance
54	13401961285604655	vvidsolution.site		ab21bf9fdc6f588232f3723e864d26e95579...

Figure 7: Cookies from “HOOHOO TV”

The .xyz domain is one of the emerging top-level domains (TLDs), notable for its extremely low registration cost (roughly $1 per year [20]) and lenient registration policies. While it is sometimes used by general users for personal websites or creative branding, it is also widely adopted for advertising delivery and tracking. Illegal streaming sites in particular have increasingly leveraged .xyz domains to track user activity and collect data for targeted advertising [21]. In practice, they frequently redirect users through .xyz-based links to hidden streaming pages or monetize traffic through pop-up ads.

According to Palo Alto Networks, .xyz ranks first in the number of grayware domains and second only to .com in the number of phishing domains [22]. Our analysis also confirms that cookies issued by .xyz domains are frequently embedded in illegal streaming sites. This reliance reflects the immaturity of their operational practices: instead of using established and trusted advertising infrastructures, these sites depend on low-cost, easily registered domains that lack accountability. Such ad hoc adoption of .xyz domains thus differentiates illegal platforms from legal services and provides a clear marker of their less professionalized operations.

5.3 Common Cookie Usage between Illegal Sites

During the artifact tracking process, we observed a high degree of consistency in the cookie and domain lists of two specific illegal streaming sites. As illustrated in Figure 8, the distribution of the top 31 duplicate cookies collected from 15 illegal Korean streaming sites further visualizes this tendency. Darker cells indicate a higher degree of cookie overlap between sites, whereas lighter shades represent lower similarity. Furthermore, the closer the observed patterns are to each other, the higher the similarity between websites. Consistent with the results from the HTML and Web cache analyses, several common cookies were again identified between “ANIWOLF” and “Aniweek”, as summarized in Table 4. A detailed comparison of all cookies and associated domain URLs from the two sites confirmed that most entries were identical. This cross-artifact consistency—spanning HTML, cache, and cookie data—strongly suggests the use of shared infrastructure or operation by the same entity, rather than a coincidental overlap.

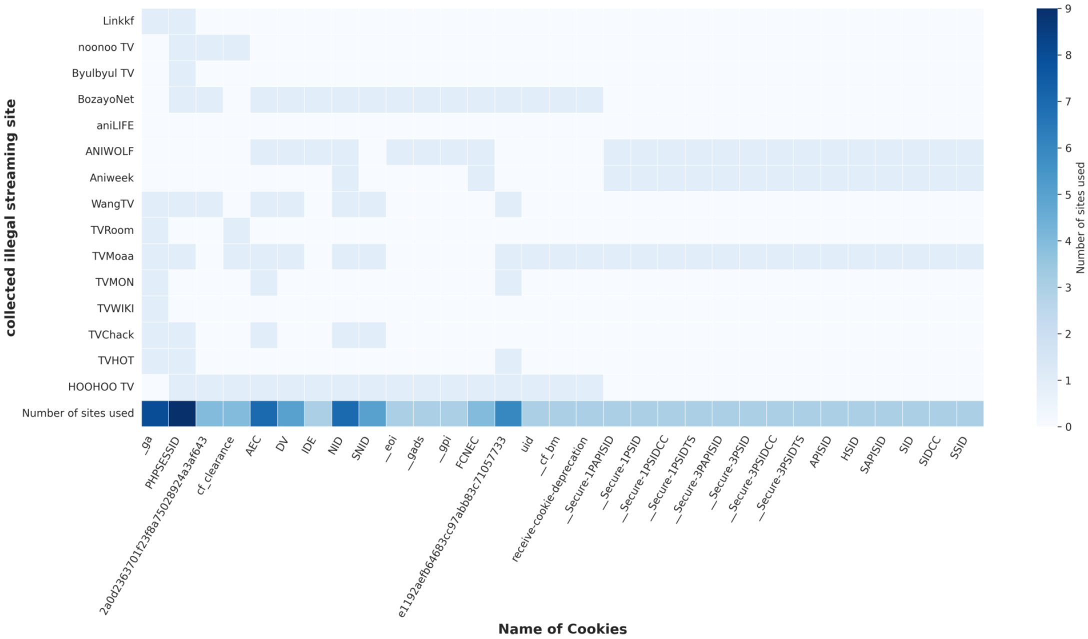


Figure 8: List of duplicate cookies extracted from 15 collected illegal streaming sites

Table 4: Common Cookies and Domain List for the "ANIWOLF" and "Aniweek" Websites

Host_key	Cookie Name
cdndania.com	fireplayer_player
.adnxs.com	anj, uuid2
.casalemedia.com	CMID, CMPRO, CMPS
.viaproducciones.net	__eoi, __gads, __gpi
.doubleclick.net	DSID, IDE

6 Conclusion

This study comprehensively compared and analyzed the HTML structure, cache, and cookies of legal streaming sites and illegal streaming sites to identify recurring technical characteristics specific to each site type. Analysis revealed that while legal streaming sites exhibit consistent resource calls based on their own domains and controlled cookie management, illegal streaming sites commonly display multiple structural characteristics. These include frequent insertion of external domains, repetitive use of specific session management cookies (e.g., PHPSESSID, cf_clearance), opaque connections with ad networks, and similar patterns of external resource calls via caching. Notably, high duplication rates in HTML, cookies, and cache entries were observed between some sites, suggesting the possibility of identical operators.

These findings not only provide technical clues for identifying illegal streaming sites but also serve as empirical data applicable to security research, digital forensic analysis, and policy responses aimed at blocking the distribution of illegal content. Future research will focus on developing a decision-making framework that integrates AI learning and automation techniques to automatically determine the legality of streaming websites, building upon the analytical results presented in this study.

Acknowledgments

This work is partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00351898 and No. RS-2025-02263915), the MSIT under the ICAN (ICT Challenge and Advanced Network of HRD) program (No. IITP-2022-RS-2022-00156310) supervised by the Institute of Information & Communication Technology Planning & Evaluation (IITP).

References

[1] Supreme Court of Korea. *Decision 2017Do19025 [Aiding and Abetting Copyright Infringement]*. Supreme Court Reports 2021 Ha, 1881, September 9, 2021.

[2] Soyoung Ham, Yeeun Kang, Seongmin Kim, Hakkyong Kim. “An Empirical Correlation Analysis of Illegal Streaming Sites Managed by the Same Operator.” *Journal of Convergence Security*, vol. 25, no. 1, pp. 93–103, 2025.

[3] Joong-won Jeong. “A Study on the Domain Change Pattern Analysis and Blocking Method of Illegal Site.” Master’s thesis, Department of Cyber Security, Korea University, 2022.

[4] Sheaib H., Feldmann, A., Dao, H., ”Unmasking the Shadows: A Cross-Country Study of Online Tracking in Illegal Movie Streaming Services” *25th Privacy Enhancing Technologies Symposium*, 2025(2), 125–139., 2025.

[5] Seungyong Choo, Yeseong Hwang, Sangjin Lee, “Methods for Collecting Harmful Websites Using Web Crawling,” *Journal of Digital Forensics*, vol. 15, no. 3, pp. 127–138, Sep. 2021.

[6] Infoblox, ”RDGAs: The Next Chapter in Domain Generation Algorithms,” Infoblox Threat Intelligence Blog, May 13, 2021. [Online]. Available: https://blogs.infoblox.com/threat-intelligence/rdgas-the-next-chapter-in-domain-generation-algorithms/.

[7] Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 93–104.

[8] Sauer, S. (2017). Effect sizes for the Mann-Whitney U Test: an intuition. Retrieved November 10, 2025

[9] Cloudflare. (n.d.). What are cookies? Retrieved September 17, 2025, from https://www.cloudflare.com/learning/privacy/what-are-cookies/

[10] CookieScript. (2022, September 6; Last Updated: December 19, 2024). What are Session Cookies and do They Need a Cookie Consent? Retrieved September 17, 2025, from https://cookie-script.com/blog/session-cookies/

[11] Cookiebot. (n.d.). What are session cookies?. Retrieved September 17, 2025, from https://www.cookiebot.com/en/session-cookies/

[12] CookieDatabase.(n.d.). PHPSESSID. Retrieved September 17, 2025, from https://cookiedatabase.org/cookie/php/phpsessid/

[13] OWASP Foundation. (n.d.). Session fixation. Retrieved September 17, 2025, from https://owasp.org/www-community/attacks/Session_fixation

[14] Cloudflare. (2025, September 4). Clearance. Retrieved September 17, 2025, from https://developers.cloudflare.com/cloudflare-challenges/concepts/clearance/

[15] “Difference Between Call by Value and Call by Reference in C.” GeeksforGeeks, Sanchhaya Education Private Limited, July 11, 2025. Retrieved September 17, 2025 from https://www.geeksforgeeks.org/c/difference-between-call-by-value-and-call-by-reference/

[16] Google. 2025. How Google uses cookies – Privacy & Terms – Google. Retrieved September 18, 2025 from https://policies.google.com/technologies/cookies?hl=en-US

[17] Lim, Joel. 2024. DV Cookie: How These Cookies Affect You. CaptainCompliance.com, May 13, 2024. Retrieved September 18, 2025 from https://captaincompliance.com/education/dv-cookie/

[18] "test_cookie – Cookie Database." Cookie.is. Retrieved September 18, 2025 from https://www.cookie.is/test_cookie#

[19] etify. (n.d.). jwpcdn.com – Domain info(JW Player). Retrieved September 17, 2025, from https://www.netify.ai/resources/domains/jwpcdn.com

[20] "Cheapest .XYZ domain registration, renewal and transfer – Compare prices of 124 registrars." TLDES, retrieved September 18, 2025 from https://tldes.com/xyz

[21] "Beware of unfamiliar addresses... Increase in phishing sites exploiting new domains," Digital Today, 2024. [Online]. Available: https://www.digitaltoday.co.kr/news/articleView.html?idxno=544331

[22] Palo Alto Networks Unit42. "A Peek into Top-Level Domains and Cybercrime." Unit42 Research, Palo Alto Networks, June 2024.