2018 Study Of The Presence Of Malicious Content At Content Theft Sites Visited By Canadian Consumers

This quantitative study is designed to analyze the amount and type of malicious content found on content theft sites against a control group of sites representing the general Internet. In both groups, a broad sample of sites were analyzed, including sites that are highly popular with Canadian consumers and those less frequently visited by Canadian consumers.

RISKIQ®

Objectives/Methodology

Objective

- MPA Canada commissioned RiskIQ to analyze the prevalence and nature of malicious content on sites that facilitate copyright infringement (content theft sites) visited by Canadian consumers. RiskIQ defines malicious content as software designed with a possible malicious intent to gain unauthorized access, collect private data, or inflict intentional damage.
- RiskIQ performed this study by analyzing the rate of malicious content exposures across a sample of content theft sites against a control group representing the general web site population.
- This study was designed to provide an objective comparison of the amount and type of malicious content found on content theft sites with a control group of sites, which comprise legitimate viewing sites and random sites that are representative of the general Internet. In both groups, a broad sample of sites was analyzed, including sites that are highly popular and those less frequently visited.

Methodology: Sample Design

- Sample Group:

  The sample group was comprised of 400 content theft sites with Canadian Alexa ranking of up to 100,000, including the Top 25 sites from the June 2018 Google Transparency Report and 350 sites selected at random from the top, middle and bottom third of the Google Transparency Report. Roughly 25 link piracy sites and French language piracy sites were also included. The sample group excluded sites primarily dedicated to video game piracy (because most gaming files are executable files, which carry more apparent inherent risks of malicious content infection) and sites with primarily adult content, which were filtered out by running a keyword classifier for page content.

- Control Group:

  The control group was comprised of 396 sites intended to represent legal online video sites and the general internet, including 60 legal online video sites available to Canadians, and 336 sites selected randomly from top, middle and bottom third of sites with a Canadian Alexa ranking of up to 100,000. The control group excluded the following types of sites: sample group/content theft sites, adult, drugs, gaming, gambling, scam, forums/blogs, non-English/French sites, and software sites designed to provide executable files.

Methodology: Data Collection

- RiskIQ scanned all 800 sites in both the sample and control group for malicious content for a period of four weeks from June 25 to July 24, 2018. The sites were probed for malicious content by simulating the behavior of users from Canada with a variety of operating systems and browsers, to approximate typical Canadian targets for malicious content distributors.
- Each simulated user was configured to navigate up to three levels deep for max of 25 pages daily per site. Data collection sampled average of 50+ pages daily per site during the four week period.
- RiskIQ's virtual users crawled links on sites and downloaded any unprotected download links (not blocked by logins or CAPTCHAs).
- Scans were designed to check for the presence of malware in either "drive-by downloads" or user-initiated downloads, typically delivered through pop-ups or fake software update requests.

Methodology: Malicious Content Analysis

- Malicious content analysis was run through a series of market standard malware detection tools, including VirusTotal and RiskIQ's own proprietary detection system.
- For this study, RiskIQ defined malicious content as software designed with a possible malicious intent to gain unauthorized access, collect private data, or inflict intentional damage.
- RiskIQ's focus is to detect malicious content incidents for websites and advertisers trying to protect themselves from malicious exploits that would affect their end consumers and partners. As such, RiskIQ detects anything ranging from suspicious incidents to outright, confirmed cases of malware. RiskIQ's system classifies cases as "exact" matches, which are confirmed and active cases, and "reputational" cases, which are incidents that are suspect because they exhibit characteristics or infrastructure that are commonly or previously associated with malicious content or behavior. These are high probability matches to malicious content. For the purposes of this study, RiskIQ reported "exact" matches and "reputational" matches separately, in order to provide the full view of data on malicious content related to these sites.

Methodology: Definitions

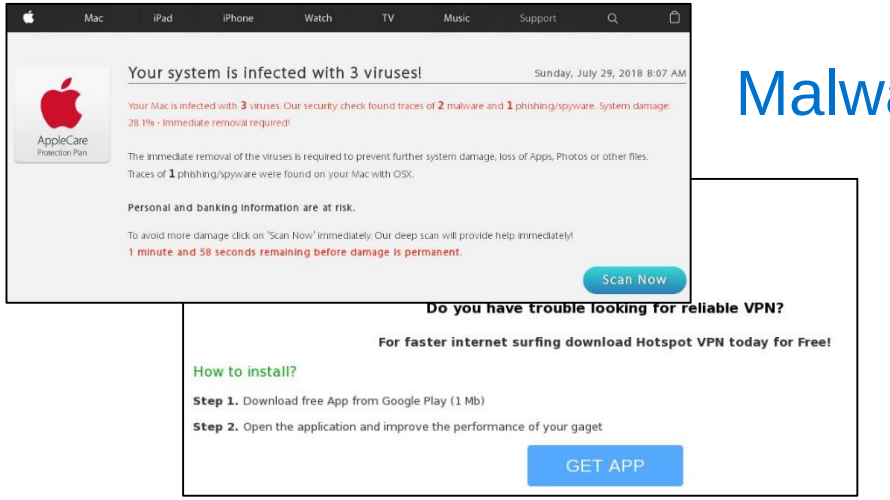
- The categories of threats to consumers seen in this study referred to as “Malicious Content” include malware, phishing, spam, scam, and malicious redirectors.
- Previous studies[1] of malicious content have focused only on malware, which used to be the majority of malicious content seen. However, the online landscape has changed such that this focus would miss other forms of malicious content that have become more prevalent in recent years. Instead of forcing installation on a user’s computer via malware, malicious content providers now often rely on other forms of malicious content that trick users into giving up their personal information. While these other types of malicious content were not highly prevalent in the past, they are included in this report to accurately convey the current landscape of malicious content delivery.

Malicious Content	
Type	**Definition**
Malware	Software designed with malicious intent to gain unauthorized access, collect private data, or inflict intentional damage. Includes Trojans, potentially unwanted programs (PUPs), adware, toolbar, botnet, and other categories
Phishing	Fake site that defrauds users to log their username and password information, often redirecting user to legitimate website afterward.
Spam	Unsolicited messages/images injected into an online advertising network and/or webpage.
Scam	A fraudulent claim to offer tech support, E-bay auctions, fake donations with goal of getting the victim to allow remote access to their computer.
Malicious Redirectors	A term that describes a domain that, while not directly serving malicious content, we have flagged as redirecting user traffic to malicious content

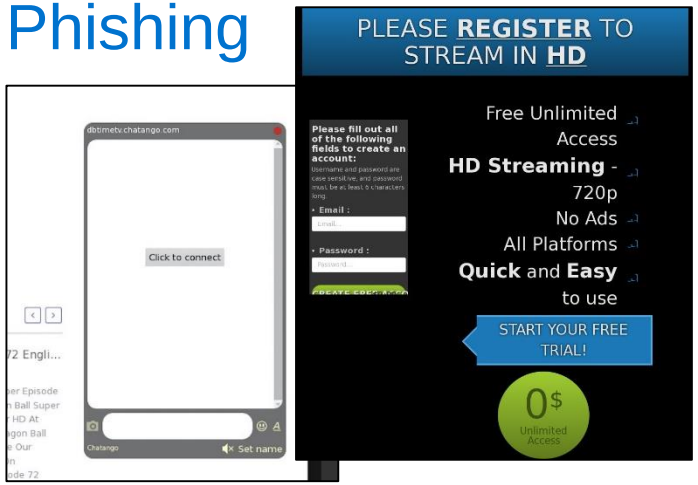
[1] https://www.digitalcitizensalliance.org/clientuploads/directory/Reports/digitalbait.pdf

Malicious Content - Examples

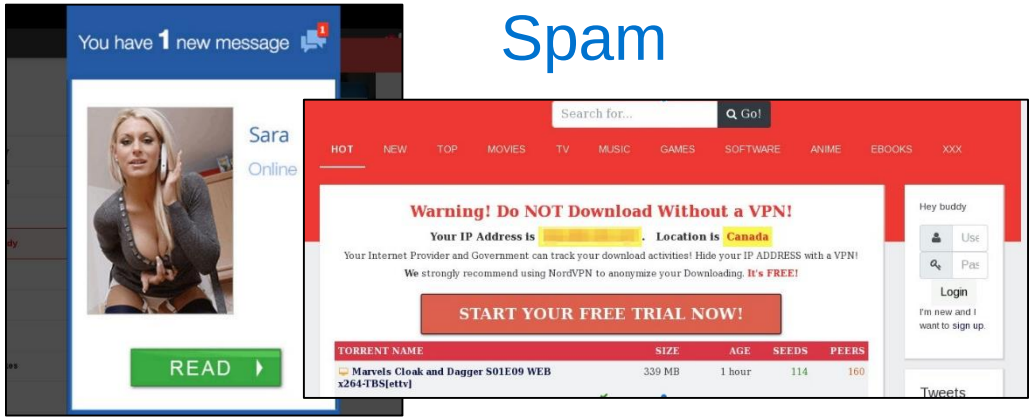
Malware



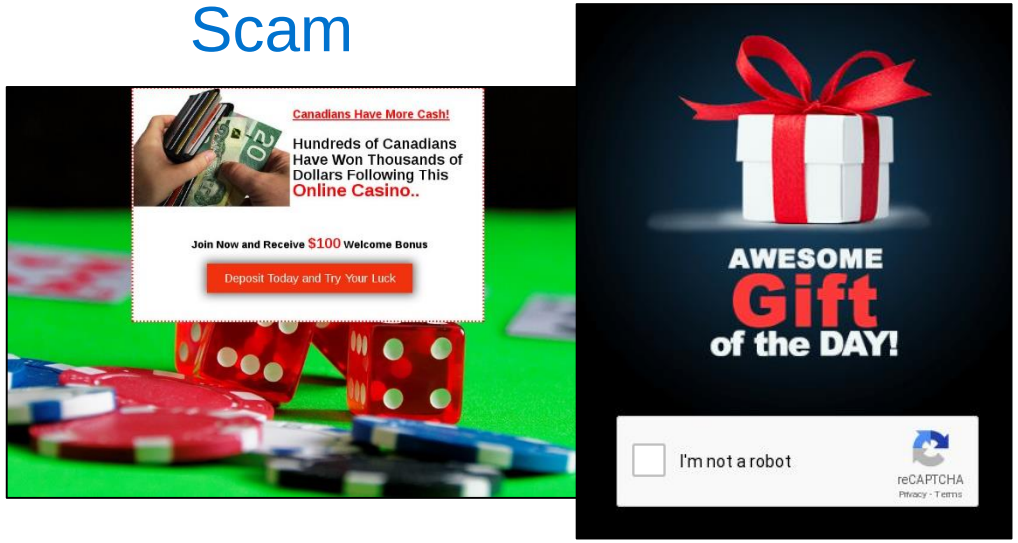
Phishing



Spam



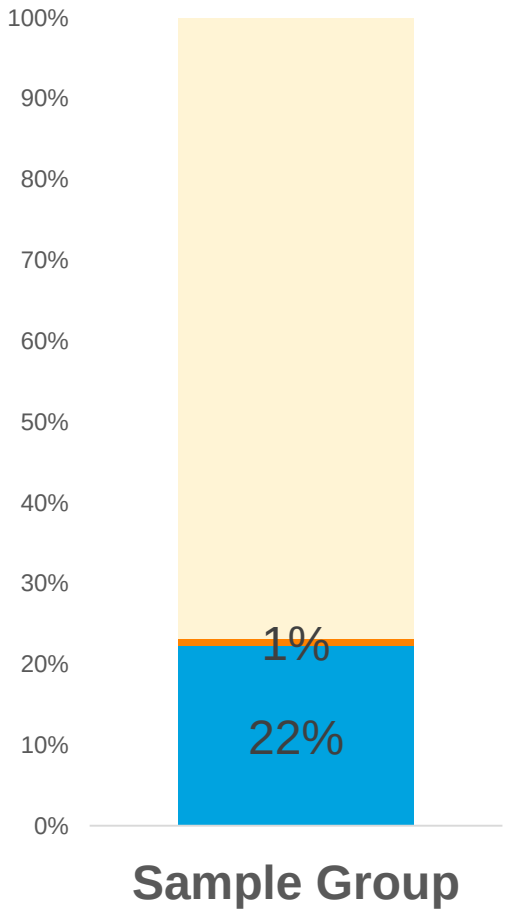
Scam



Findings

RISKIQ®

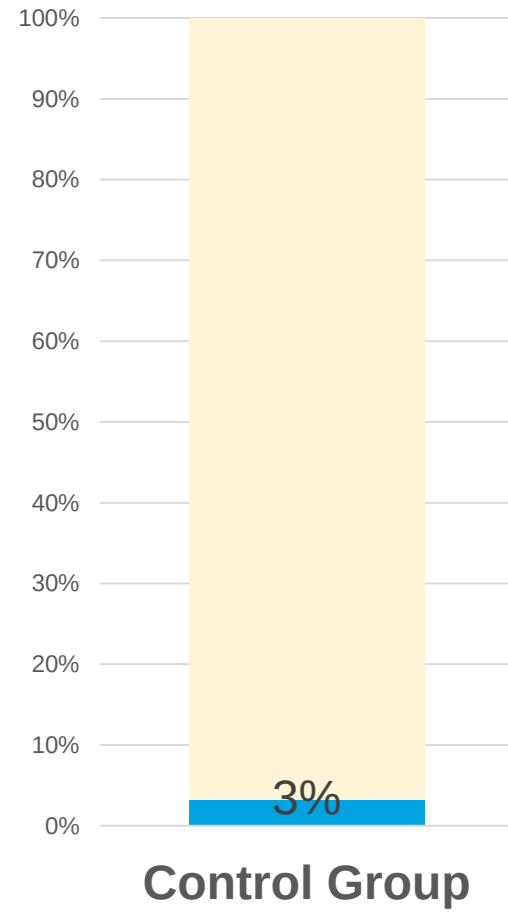
Rate of Presence of Malicious Content by Domain Type



23% of content theft sites (sample group) exposed consumers to malicious content (malware, phishing, spam, scam or malicious redirectors) compared to 3% of sites in the control group.

22% of the content theft sites were flagged for malicious content based on exact matches, while an additional 1% of content theft sites were flagged based on reputational matches, or high probability matches based infrastructure and prior behavior.

In other words, nearly 1 in 4 content theft sites distributed malware, phishing, spam, or scam content to consumer, or redirected them to such content, versus 1 in 30 control group sites.[2]

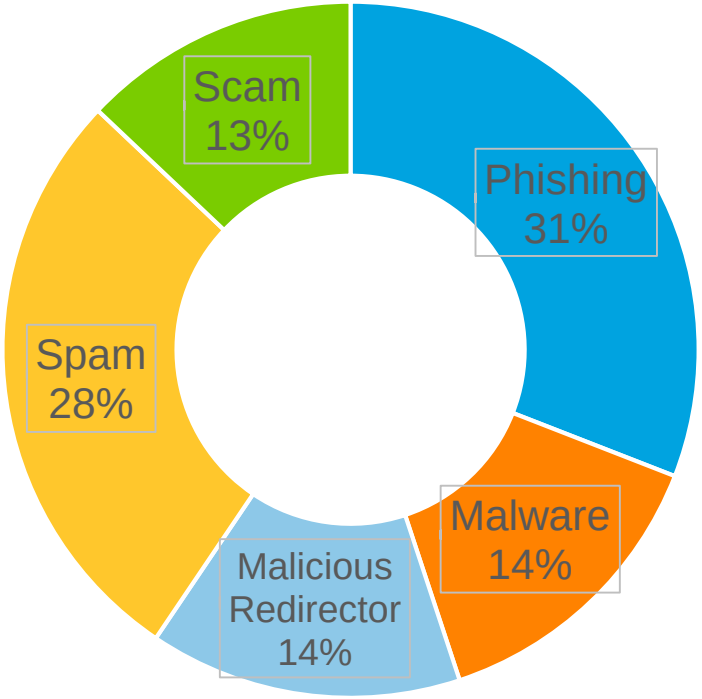


[2]For the control group sample, note that one site (soundkeepers.com) makes up 88% of all malicious content incidents.

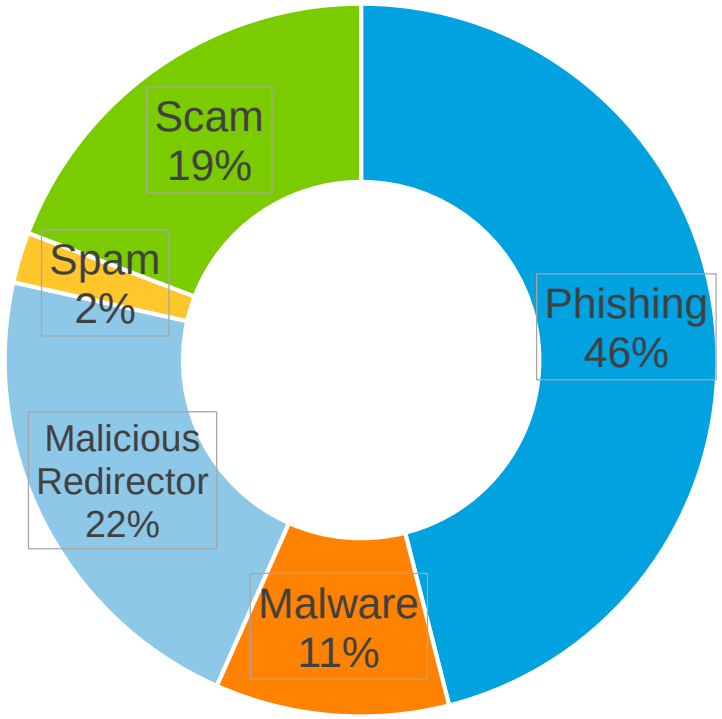
Malicious Content on Content Theft Sites by Category

- Phishing is the most prevalent forms of malicious content within the content theft sample, at 31% of the malicious content encountered overall, and 46% of the exact matches. These types of malicious content are meant to get users to give up their personal information.
- Spam (28%), Malware (14%), Malicious Redirectors (14%) and Scam (13%) are also prevalent. Spam is primarily high probability matches based on infrastructure/behavior (reputational) so there are fewer exact matches for Spam. The other categories' share increases when the focus is on exact matches.

Content Theft Sample: All Matches



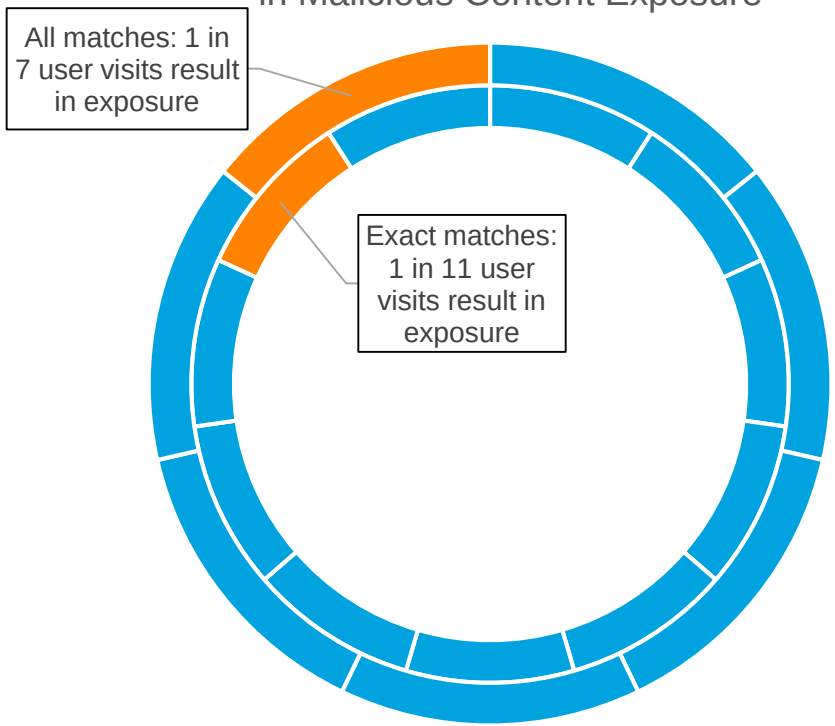
Content Theft Sample: Exact Matches Only



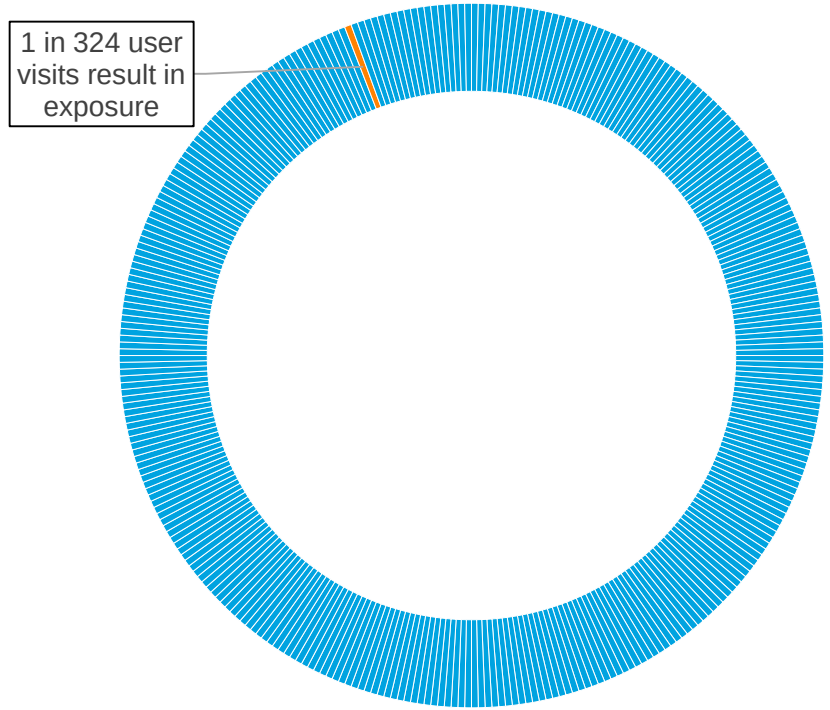
Malicious Content Incident Rates Compared to Control

- Users are 46 times more likely to be exposed to malicious content on a content theft site than on a site in the control group, when considering all matches to malicious content – and 30 times more likely to be exposed to malicious content on a content theft site when only exact matches are included.
- 15% of user visits to content theft sites resulted in exposure to malicious content, compared to 0.3% of user visits for the control group. 9% of content theft site visits were exposed based on an exact match to malicious content, with the remainder being high probability reputational matches.

Share of Content Theft Sites' User Visits Resulting in Malicious Content Exposure



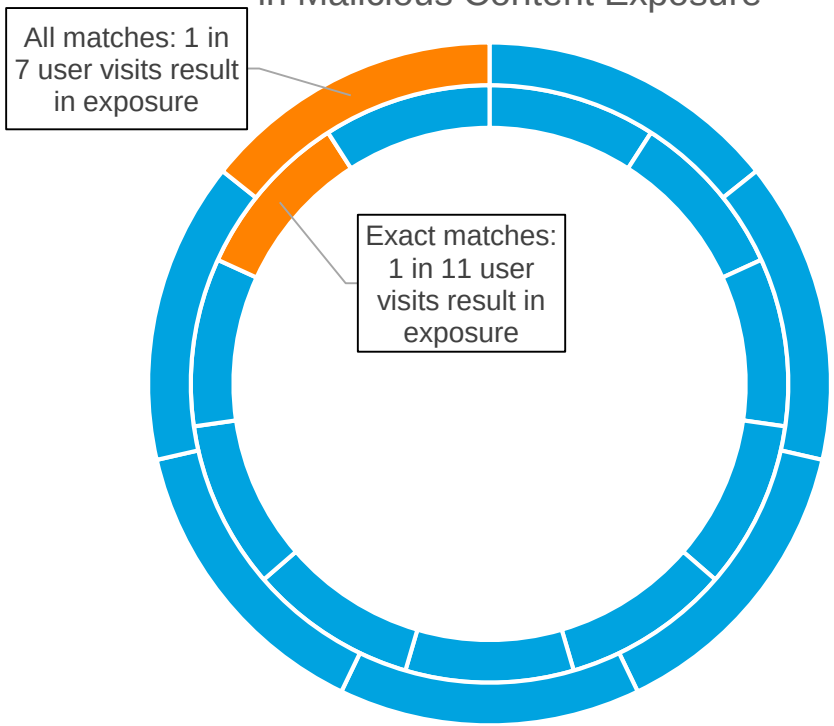
Share of Control Group Sites' Users Visits Resulting in Malicious Content Exposure



Malicious Content Incident Rates Compared to Online Video

- Users are 100% more likely to be exposed to malicious content on a content theft site than on a legal online video site in the control group, as 1 in 7 of user visits to content theft sites resulted in exposure to malicious content based on all matches, compared to none of the user visits for legal online video sites in the control group.

Share of Content Theft Sites' User Visits Resulting in Malicious Content Exposure



Share of Legal Online Video Sites' User Visits Resulting in Malicious Content Exposure

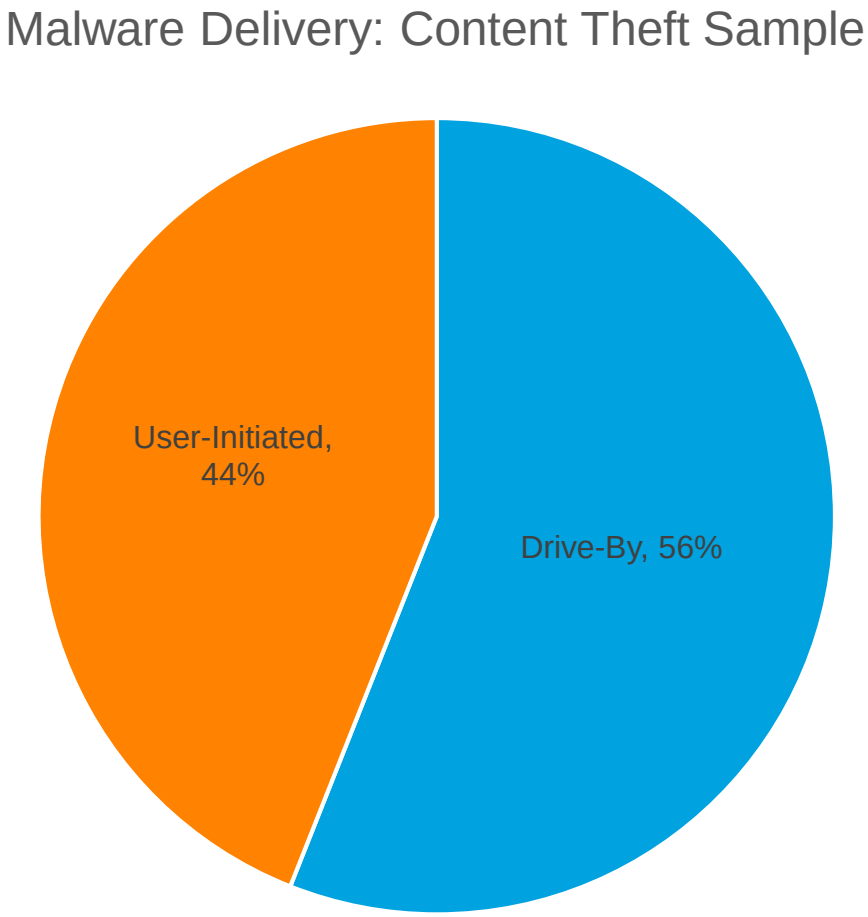
None

Malware Delivery

- 14% of all malicious content encountered on content theft sites was malware, based on all matches . Malware downloads malicious content onto a user's computer, either via a "drive-by download" or via a "user-initiated download."

- Drive-by downloads allow malware to be delivered without users having to do anything to confirm the download. Drive-by-downloads infect more users because the users don't have to click on them.

- User-initiated downloads lure users with fake prompts that they click to allow the download. User initiated downloads use deceptive links to trick users but may have bigger payloads with more malware.
-
- In the content theft sample, 56% of all malware were drive-by, and 44% were user initiated.



Conclusions And Recommendations

- Content theft is not just a copyright issue; it supports a criminal ecosystem of malicious content including malvertising and malware distribution that harms the browsing public
- More attention and resources need to be devoted to apprehending these criminals
- Digital platforms and financial facilitators need to ensure they are not aiding cyber criminals