Detecting and Blocking Site Scraping Attacks

Protecting Valuable Data and Intellectual Property from Online Thieves

Data theft is one of the most dangerous threats for online businesses today. And site scraping tools have become the weapon of choice to perform data theft. From highly publicized scraping attacks on sites like Google, RyanAir, and EasyJet, to innumerable unpublished incidents, scraping has become a wide-spread menace.

Scraping attacks range from harmless data collection for personal research purposes to calculated, repeated data harvesting used to undercut competitor's prices or to illicitly publish valuable information. Site scraping can undermine victims' revenues and profits by siphoning off customers and reducing competitiveness. Unfortunately, site scrapers can be very difficult to differentiate from legitimate Web application users. This paper investigates various types of scraping attacks, site scraping tools, and effective techniques to detect and stop future attacks.

Introduction

When you picture a thief, do you envision a masked man lurking in a dark alley? Or a teenage hacker in a dimly-lit basement breaking into remote Web servers? What about car manufacturers and online travel, software, financial news, and event planning companies? All of these organizations and many others have been accused of stealing data by site scraping.

Site scraping, alternately called data harvesting, screen scraping or Web scraping, is a technique used to extract data from a Website. Site scraping can be performed by manually copying Website content from Web browsers. However, the overwhelming majority of site scraping attacks are performed by software programs that automatically extract data from Websites. Site scraping tools range from simple, home grown scripts to commercial scraping software with embedded Web browsers that can parse HTML and store scraped data in databases.

Site scraping can be used for legitimate purposes, such as search engine indexing or mashups like comparative shopping tools. Site scraping can also offer a way for individual users to quickly and automatically gather data, such as stock price information or daily temperatures, which can be used for personal research. While not blatantly illegal, such actions can burden sites with traffic while bypassing online ads.

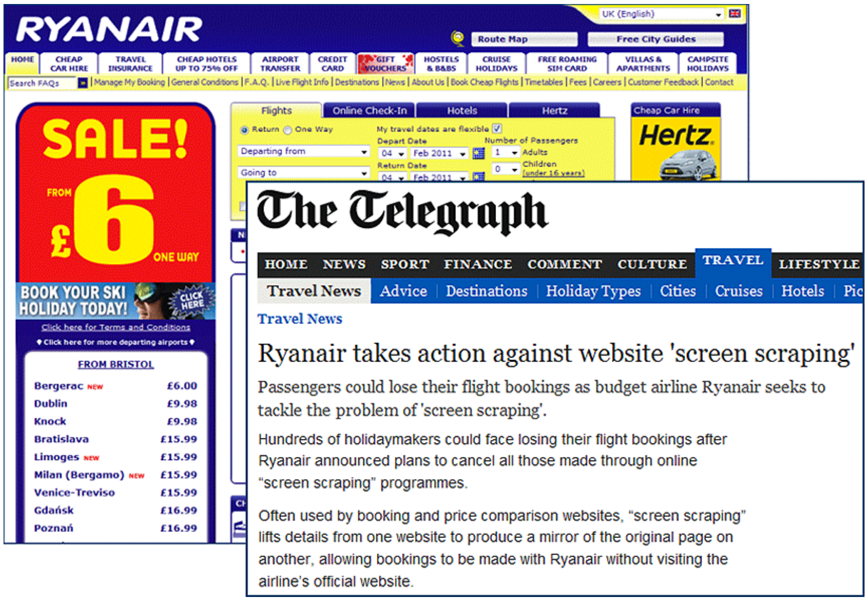
Site scraping quickly crosses over from being innocuous to malicious when individuals or businesses reuse scraped data for financial gain. The following are just a few of the use cases for site scraping:

- Collecting email addresses from Websites for spam email campaigns
- Harvesting user information from social network sites or user forums
- Detecting changes on competitors' Websites
- Gathering product and pricing data to undercut rivals' prices for goods and services
- Plagiarizing content such as news articles, blog posts, medical information, financial research
- Republishing Website listings, such as job board postings, real estate listings, and phone directories
- Repurpose scraped content for applications such as comparison shopping sites or reverse phone lookup tools
- "Auction sniping" or placing bids on online auction sites within minutes or seconds of the auction ending

The actions listed above violate the intended use for application data. For Web-based companies, site scraping threatens revenues and competitiveness and can even endanger business viability.

High Profile Site Scraping Attacks

While the vast majority of scraping attacks are not publicized, several incidents have garnered national and even international attention. Discount Irish airline Ryanair has embarked on a multi-year battle against companies that it believes are scraping its Website. Ryanair sent cease and desist notices to over 300 travel reservation and comparison sites in 2008, prohibiting them from scraping the Ryanair Website. Ryanair subsequently sued companies in Germany, Spain, The Netherlands, and Ireland for scraping flight data and then selling airline tickets to customers at inflated prices.



FROM BRISTOL

Bergerac NEW	£6.00
Dublin	£9.98
Knock	£9.98
Bratislava	£15.99
Limoges NEW	£15.99
Milan (Bergamo) NEW	£15.99
Venice-Treviso	£15.99
Gdańsk	£16.99
Poznań	£16.99

Travel News

Ryanair takes action against website 'screen scraping'

Passengers could lose their flight bookings as budget airline Ryanair seeks to tackle the problem of 'screen scraping'.

Hundreds of holidaymakers could face losing their flight bookings after Ryanair announced plans to cancel all those made through online "screen scraping" programmes.

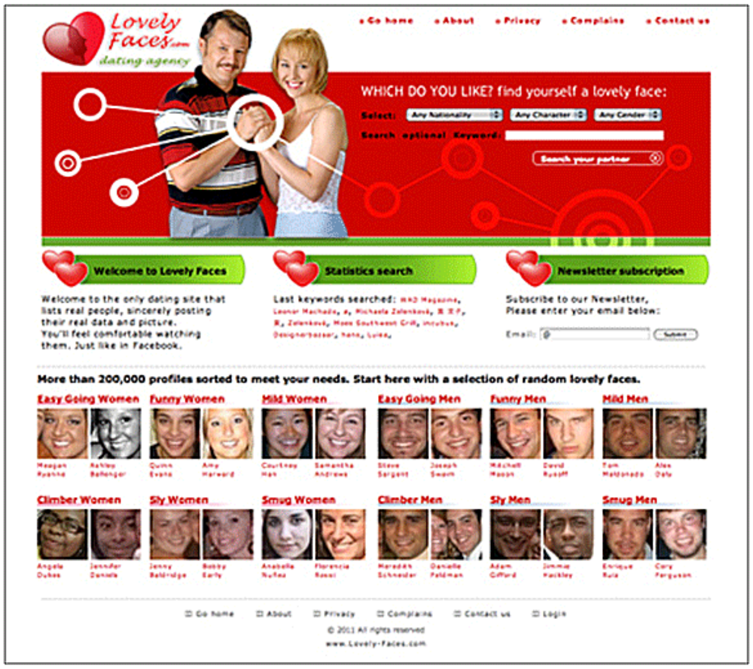
Often used by booking and price comparison websites, "screen scraping" lifts details from one website to produce a mirror of the original page on another, allowing bookings to be made with Ryanair without visiting the airline's official website.

Ryanair's legal action against site scrapers resulted in multiple lawsuits and the attempted cancellation of flights booked through third party sites.

Ryanair is not alone; several other airlines including American Airlines and Southwest Airlines have sued travel sites for their site scraping practices. Southwest Airlines claimed that scraping by FareChase and Outtask constituted "Computer Fraud and Abuse" and "Trespassing" and led to "Damage and Loss" for the carrier. Although the cases were never resolved in the Supreme Court, FareChase was eventually shuttered by parent company Yahoo! and Outtask was purchased by travel expense company Concur.

In another case, event planning Website Cvent sued rival Eventbrite for hiring a third party contractor to scrape Cvent's Website for event venue information. Eventbrite subsequently republished the scraped content on its own site. In October 2010, Cvent and Eventbrite settled the litigation, with Eventbrite agreeing that it would limit the scope of its venue listings and it would not scrape nor hire third parties to scrape the Cvent site.

Many other companies, including eBay and Facebook, have taken legal actions against scrapers that harvest user information and product listings. An eBay lawsuit against Bidder's Edge resulted in an injunction preventing Bidder's Edge from harvesting data from the eBay Website. FaceBook has sued several pornography and dating sites that had scraped FaceBook user profiles. One such site, Lovely-Faces.com, displayed 250,000 user profiles that had been scraped from Facebook's Website. The Lovely-Faces.com creators used an automated "Facebot" to scrape one million profiles to a Face-to-Facebook database. Then Facebook profile photos were analyzed by facial recognition software to classify users into categories such as "easy going," "smug," "climber," "sly," and "funny." Facebook users were then showcased on the Lovely-Faces.com Website.



Lovely Faces.com dating agency

Go home · About · Privacy · Complains · Contact us

WHICH DO YOU LIKE? find yourself a lovely face:

Select: Any Nationality · Any Character · Any Gender

Search optional Keyword:

Search your partner

Welcome to Lovely Faces

Welcome to the only dating site that lists real people, sincerely posting their real data and picture.
You'll feel comfortable watching them. Just like in Facebook.

Statistics search

Last keywords searched: [illegible], Leonor Machado, a, Michaela Zelenková, [illegible], [illegible], Zelenková, [illegible], [illegible], [illegible], [illegible], [illegible],

Newsletter subscription

Subscribe to our Newsletter,
Please enter your email below:

Email: @ Submit

More than 200,000 profiles sorted to meet your needs. Start here with a selection of random lovely faces.

Easy Going Women: Meagan Ryanne, Ashley Bellenger
Funny Women: Quinn Evans, Amy Harward
Mild Women: Courtney Han, Samantha Andrews
Easy Going Men: Steve Sargent, Joseph Swam
Funny Men: Mitchell Mason, David Ryoff
Mild Men: Tom Maldonado, Alex Dale

Climber Women: Angela Dukes, Jennifer Daniels
Sly Women: Jenny Baldridge, Bobby Early
Smug Women: Anabelle Nuñez, Florencia Rossi
Climber Men: Meredith Schneider, Danielle Feldman
Sly Men: Adam Gifford, Jimmie Hackley
Smug Men: Enrique Ruiz, Cory Ferguson

□ Go home □ About □ Privacy □ Complains □ Contact us □ Login

© 2011 All rights reserved
www.Lovely-Faces.com

Debuting in February 2011, the Lovely-Faces.com Website showcased hundreds of thousands of scraped Facebook user profiles. Lovely-Faces.com was quickly taken down after Facebook threatened legal action.

Recent, high-profile incidents represent a small fraction of all site scraping attacks. Site scraping has become a pervasive challenge for hundreds of thousands of Web-based businesses.

Scraping Tools and Techniques

Dozens of vendors advertise site scraping solutions that "quickly and efficiently harvest information," while saving clients "hundreds of thousands of man-hours and money." Because site scraping software can be easily developed, many different tools have sprung up to scrape Web content for a multitude of purposes. Scraping vendors brazenly claim that their solutions can gather sales leads, capture job postings, harvest product pricing data, collect dating site information, and duplicate online databases. Such use cases are often illegal because they violate the Terms of Use policies of the targeted sites.

Site scraping software ranges from simple, custom scripts to advanced software tools with built-in browser-like capabilities to parse HTML and DOM (Document Object Model) and to interpret JavaScript. Some scraping tools have been designed to impersonate normal users—they provide Internet Explorer or Mozilla Firefox browser agent information, they can login to privileged areas of a Website, and they can limit the rate of requests to simulate manual Web browsing. Many scraping tools also include Web site crawlers that help automate initial configuration.

Imperva's research organization, the Application Defense Center (ADC), has observed a multitude of site scraping attacks against online phone directories, finance sites, job postings sites, and many more. Analyzing real application traffic, including Tor honeypot logs, the ADC has uncovered a variety of site scraping attacks. While some advanced attacks leveraged scraping software such as AutoHotkey, others relied on homegrown scripts. In one case, a Microsoft Excel spreadsheet distributed by eTrader Zone was used to scrape historical stock quotes from the Yahoo! Finance. While not illegal or malicious, this tool could potentially violate the site's Terms of Use policy by exceeding reasonable request volumes.

Site Scraping Mitigation Strategies

There are several ways that businesses can reduce or eliminate site scraping: take legal action against site scrapers, implement barriers in the Web application to hinder scraping attacks, and block scraping clients using a Web Application Firewall.

Since there are no clear laws that prevent the harvesting of public data, using legal channels to stop site scraping can be challenging and expensive and require years to resolve. Victims must clearly demonstrate that site scrapers have violated the Terms of Use policy for the Website. Victims have legally prohibited site scraping by proving that defendants committed "Trespass to Chattels" for misusing personal property or performed "Computer Fraud and Abuse" or "Unauthorized Access." While legal decisions in scraping cases have been mixed, many defendants stop scraping victims' sites in order to avoid further legal action.

Alternatively, organizations can prevent site scraping activity by building anti-automation obstacles into their applications. There are options like CAPTCHAs and code obfuscation that can prevent script-based scrapers from harvesting content. However, more advanced scraping tools can evolve to circumvent such measures, rendering code-based mitigation techniques ineffective over the long term.

Lastly, organizations can deploy a Web Application Firewall to proactively detect and block scraping attacks. Since site scrapers can be extremely difficult to differentiate from legitimate users, identifying scrapers can be extremely challenging. While simple scraping tools are relatively easy to detect, advanced scraping software may require a combination of detection methods.

The following are a list of security measures that can be used by a Web Application Firewall or implemented via code changes to prevent site scraping attacks:

- **Use cookies or JavaScript to verify that the client program is a standard Web browser**—Most simple scraping tools cannot process complex JavaScript code or store cookies. To verify that the client is a real Web browser, inject a complex JavaScript calculation and determine if JavaScript is correctly computed.
- **Implement CAPTCHAs to verify that the user is human**—CAPTCHAs[1] can significantly hinder site scraping attacks. However, bots are increasingly finding ways to circumvent CAPTCHAs. Up to 60 percent of bots can crash through CAPTCHAs, according to recent security research.[2] Nevertheless, CAPTCHAs are still an effective defense against many types of scraping agents.
- **Rate limit requests**—Automated clients typically request Web pages much more frequently than real users. Advanced scraping clients will try to evade detection by slowing down scraping request rates. However, automated clients will expose themselves by requesting Web content after a consistent time period. Scrapers are also more likely to open an unusually high number of concurrent connections and access many more pages than most users. Some scrapers will only access html files and ignore images and other binary files.
- **Obfuscate data**—Site scrapers are designed to extract text from Web pages. Displaying Web content as images or flash files can hinder site scrapers. Alternatively, because most scraping tools cannot interpret JavaScript or CSS, applications can compile text using script or style sheets to evade scraping tools.
- **Detect and block known malicious sources**—Since competitors often perpetrate scraping attacks, blocking competitors' IP address ranges can deter scraping activity. Many site scrapers use evasion techniques to conceal their identity such as anonymous proxies, Tor network servers, or leased botnets. Detecting and blocking known malicious sources can hinder scraping attacks.

[1] CAPTCHAs are tests to distinguish between computers and humans. A common CAPTCHA requires that users type characters from a distorted online image.

[2] "Botnets Target Websites with 'Posers,'" Dark Reading, June 1, 2010.

- **Detect and block known bot agents and scraping tools**—Most scraping activity is performed using simple tools that have identifiable signatures, such as a unique user agent string. Tools that recognize bot agents can immediately stop these scraping tools.
- **Constantly change html tags to deter repeated scraping attacks**—Scrapers must be programmed to parse sites and extract desired data. Altering HTML tags and Web structure by, for example, adding white spaces and comments, changing tag names, hyperlink strings, and URLs can impede repeat site scraping attacks.
- **Use fake Web content, images, or links to ensnare site scrapers**—If an organization suspects that its proprietary data is being plagiarized, it can produce fictitious content as proof. Monitoring services purposely designed to identify plagiarism or even a simple online search can uncover fictitious content published on other sites. Since some scrapers include an automated crawler, creating a hidden hyperlink to a fake page can trap an automated scraper.

Since several of the detection techniques are not decisive, they can be used in combination to accurately pinpoint site scraping attacks. For example, excessive Web requests may indicate a scraping attack or it may reveal an AOL proxy address representing hundreds of simultaneous users. Some legitimate users disable cookies and JavaScript. However, extremely fast request rates detected together with browser irregularities often indicate an automated attack such as site scraping. If organizations implement the mitigation techniques listed above, they can significantly reduce or eliminate site scraping.

Practical Measures to Reduce Site Scraping Attacks

Site scraping attacks can reduce company revenues and profits by stealing Web site visitors and copying valuable content. Scraping can also undermine business competitiveness because rivals can use scraped pricing information to undercut product prices. Therefore, site scraping attacks can cost organizations millions of dollars in lost business.

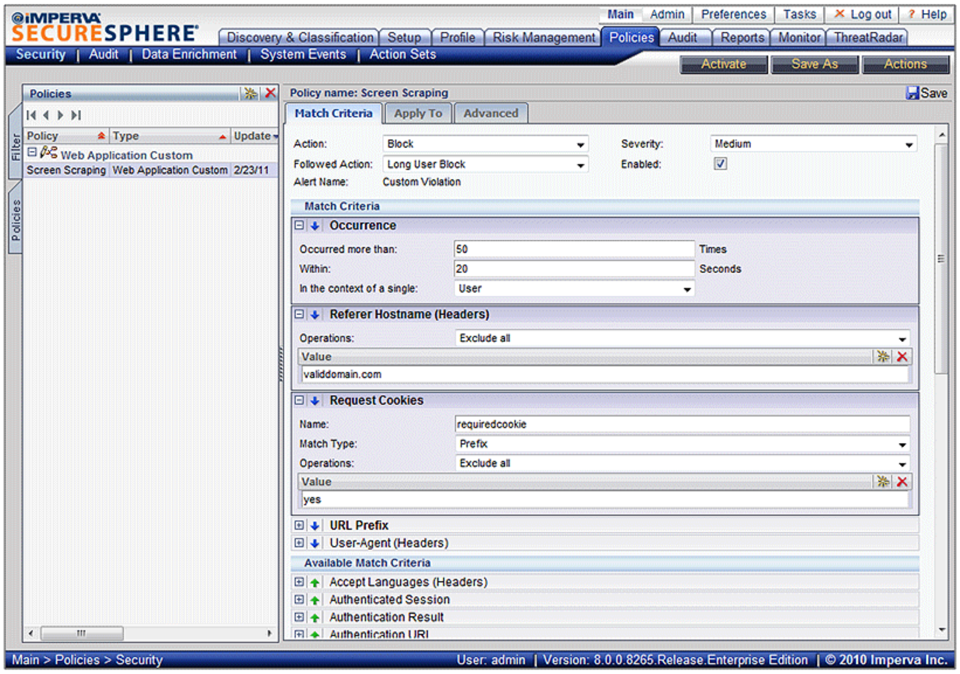
To mitigate site scraping activity, victims can take legal action against perpetrators, they can implement barriers in their applications to stop automated users, or they can proactively block site scraping exploits. The Imperva SecureSphere Web Application Firewall provides an effective defense to eliminate site scraping attacks in real-time. SecureSphere offers unique detection techniques that can identify and stop automated attacks like site scraping. In addition, SecureSphere offers flexible customization, allowing organizations to fine tune security rules based on application-specific requirements.

SecureSphere Protection against Site Scraping

Imperva SecureSphere offers multiple layers of protection to identify and stop site scraping. SecureSphere fortifies Web applications using:

- **Protection against known malicious sources**—Many site scrapers use anonymous proxies or Tor networks to conceal their identity. Site scrapers also leverage botnets for concealment or for large-scale scraping attacks. Imperva's industry-first reputation-based security service detects and optionally blocks anonymous proxies, Tor networks, and known malicious IP addresses. ThreatRadar receives near real-time feeds of known bad users from security research organizations located around the world.
- **Bot agent detection**—Site scraping is typically performed by dedicated scraping software. SecureSphere can identify and stop hundreds of the most common bot agents that are used to perform site scraping.
- **Cookie enforcement**—Most site scraping agents can imitate a legitimate Web browser, but they do not support all browser capabilities, such as storing cookies set by a Web server using JavaScript. SecureSphere can require the existence of these cookies to verify that the client program is a true Web browser.
- **Forceful browsing**—SecureSphere can be configured to detect forceful browsing, when users do not access Web pages in an expected order. SecureSphere can confirm that the Web pages referers are from the same site. Users can create custom rules to enforce specific referers per URL, enforcing the specific order of Web page requests.
- **Rate limiting**—Typically, site scraping agents request Web pages much more quickly than a human user. Site scrapers also usually access—and harvest—many more pages than average users. SecureSphere can be configured to detect and block excessive requests.

- **Custom security rules**—Because site scraping can be difficult to distinguish from legitimate requests, analyzing multiple attributes together will more accurately identify site scraping. Custom Web security rules can be defined that will block a user that requests dozens of pages in a short period of time, does not have a valid session token and does not have a correct Web page referrer. Custom rules can be built using over two dozen match criteria including HTTP protocol violations, number of occurrences, user agents, signatures, referrer URL, request header, and cookie data. Using SecureSphere's powerful custom security rules, organizations can accurately block site scrapers with an extremely low rate of false positives.


IMPERVA SECURESPHERE
Main | Admin | Preferences | Tasks | Log out | Help
Discovery & Classification | Setup | Profile | Risk Management | Policies | Audit | Reports | Monitor | ThreatRadar
Security | Audit | Data Enrichment | System Events | Action Sets
Activate | Save As | Actions
Filter
Policies
Save
Match Criteria | Apply To | Advanced

Policy	Type	Update
Web Application Custom		
Screen Scraping	Web Application Custom	2/23/11

Policy name: Screen Scraping

Action: Block
Followed Action: Long User Block
Alert Name: Custom Violation
Severity: Medium
Enabled: ☑

Match Criteria

Occurrence
Occurred more than: 50 Times
Within: 20 Seconds
In the context of a single: User

Referer Hostname (Headers)
Operations: Exclude all
Value
validdomain.com

Request Cookies
Name: requiredcookie
Match Type: Prefix
Operations: Exclude all
Value
yes

URL Prefix
User-Agent (Headers)

Available Match Criteria
Accept Languages (Headers)
Authenticated Session
Authentication Result
Authentication URL

Main > Policies > Security
User: admin | Version: 8.0.0.8265.Release.Enterprise Edition | © 2010 Imperva Inc.

Configuring a custom site scraping policy in SecureSphere

- **Real-time monitoring and analytics**—SecureSphere offers detailed security alerts that enable customers to monitor scraping attacks and suspicious events. For example, if alerts reveal suspected scraping activity from a competitor's IP address, a customer can create a rule to permanently block that IP address. Security alerts include the entire Web request, the source address, the time of day, type and severity of the alert, and a link to the policy that triggered the violation. Clear, complete alerts provide customers full visibility into site scraping activity.

Building a Strategy to Stop Site Scraping Attacks

Site scraping attacks target a myriad of organizations, including online retailers, auction sites, airline, real estate, social networking, online job listing companies, and many more. Site scraping attacks can cost organizations millions of dollars because perpetrators can repost content, steal customers, and undercut product prices.

Mitigating site scraping attacks is difficult because site scrapers are hard to distinguish from legitimate users. Therefore, an effective solution must evaluate multiple attributes to correctly identify site scrapers. SecureSphere offers ironclad protection against automated attacks like site scraping because it detects malicious source IP addresses, bot user agents, session information, and other unusual activity. Detailed security alerts and a powerful graphical reporting engine make it easy for organizations to monitor site scraping activity and measure mitigation efforts.

Organizations can rely on Imperva SecureSphere to stop site scraping attacks. The market-leading SecureSphere Web Application Firewall offers protection against a myriad of Web application threats, including SQL injection, XSS, CSRF, directory traversal, site reconnaissance, sensitive data leakage, message board comment spam, and more. Imperva SecureSphere is trusted by organizations around the world to stop Web application security threats like site scraping.

About Imperva

Imperva, pioneering the third pillar of enterprise security, fills the gaps in endpoint and network security by directly protecting high-value applications and data assets in physical and virtual data centers. With an integrated security platform built specifically for modern threats, Imperva data center security provides the visibility and control needed to neutralize attack, theft, and fraud from inside and outside the organization, mitigate risk, and streamline compliance.

www.imperva.com

© Copyright 2014, Imperva
All rights reserved. Imperva and SecureSphere are registered trademarks of Imperva.
All other brand or product names are trademarks or registered trademarks of their respective holders. WP-SITE-SCRAPING-ATTACKS-0314.1