Data Scraping On The Internet (Web Scraping)

Introduction

Web scraping, or data scraping, is an operation used for extracting data from websites. While web scraping can be done manually, the term typically refers to automated processes implemented using a web crawler or bot. It is a form of electronic copying, in which data is gathered and copied from the web for later analysis and use.

There are methods that some websites use to prevent web scraping, such as detecting and disallowing bots from crawling their pages. In response, there are web scraping systems that rely on using techniques in DOM parsing, computer vision and natural language processing to simulate human browsing to enable gathering web page content for offline parsing.

Potential Legal Violations Of Data Scraping

In order to evaluate the risks of a data scraping business model, it is essential to recognize the potential legal violations that might transpire.

Computer Fraud and Abuse Act (CFAA)

The CFAA is a federal statute that imposes liability on someone who "intentionally accesses a computer without authorization or exceeds authorized access, and thereby obtains…information from any protected computer." A determination of liability will typically focus on whether the data scraper has knowledge that the terms governing access to the website prohibit the data scraping activity.

Breach of Contract

If a user is bound by terms of service that clearly prohibit data scraping, and a user violates such terms, such a breach can be the basis for prohibiting the user's access and ability to scrape data. Whether or not such a breach of contract would result in liability to the user depends upon whether the website operator can establish that it incurred damages as a result of the breach.

Copyright Infringement

If the content obtained through data scraping is protected by copyright then the data scraper will have committed copyright infringement.

Not all data on a website is copyrightable, or if copyrightable, is not necessarily owned by the website. Many data heavy websites are composed of user-generated and owned content (such as Facebook, Twitter, LinkedIn, etc.).

Additionally, much content that is scraped is factual data, which is generally not protectable under copyright law. However, compilations of facts are treated differently, and may be copyrightable material.

The Supreme Court decision in **Feist Publications, Inc. v. Rural Telephone Service Co.** clarified the requirements for copyright in compilations. The Feist case denied copyright protection to a "white pages" phone book. In making this ruling, the Supreme Court held that copyright protection requires creativity, and no amount of "hard work" can transform a non-creative alphabetical list of phone numbers into copyrightable subject matter.

Trespass to Chattels

Trespass to chattels is a tort when a party intentionally interferes with another's possession of a property. A website owner has an enforceable property right in the servers hosting the website, so unauthorized access could constitute this tort. U.S. courts have acknowledged that users of "scrapers" or "robots" may be held liable for committing trespass to chattels which involves a computer system itself being considered personal property upon which the web scraper is trespassing. Courts have found such a trespass mostly where the scraping puts a burden on the website operation.

Recent Court Cases

1. **Craigslist Inc. v. 3Taps Inc.**

3Taps and PadMapper are companies that partnered to provide an alternative user interface for real estate advertisements, filled with data scraped from Craigslist's site. Craigslist then sued, resulting in this case which was litigated in the District Court for the Northern District of California.

3Taps claimed that it had Craigslist's authorization to access the data, and that Craigslist was a public website; therefore anyone, including 3Taps, was authorized. The court disagreed with this, stating that Craigslist had initially granted 3Taps authorization, however it then revoked that authorization through a subsequent cease-and-desist letter and IP blocking.

3Taps stated that an ordinary user would be more likely to misunderstand Craigslist's IP blocking than, for example, a system that required a password to gain access. The court found this not of much concern, highlighting that the personalized cease-and-desist letter and subsequent lawsuit provided adequate notice and information. The court found that this would be sufficient in differentiating the case from more benign incidents where a user accidentally stumbles upon a protected system. The court admitted that it could not comment on whether it would consider Craigslist's IP blocking to be effective, but considered the fact that 3Taps went out of its way to bypass it as enough evidence that 3Taps acted without authorization.

2. **QVC Inc. v. Resulty LLC**

QVC is a well-known TV retailer. Resultly is a shopping app which builds a catalog of items for sale by scraping online retailers, including QVC. In May 2014, Resultly began scraping the QVC website. QVC's

website's terms of use didn't prohibit scraping. Soon after Resultly began scraping, QVC's servers experienced an overload that prevented consumers from making purchases on the QVC website and resulted in an alleged $2 million loss to QVC. QVC claimed that the overload was caused by the speed at which Resultly scraped its server and sued Resultly under the CFAA.

In QVC Inc. v. Resulty LLC, the Pennsylvania district court considered whether a scraper violated the CFAA's prohibition on knowingly causing the transmission of code and intentionally causing damage, without authorization, to a protected computer. The central question was whether Resultly intended to cause damage to QVC when it scraped its website.

The court found that in order to prove that a defendant intended to cause damage to a computer, the evidence had to show that it was the defendant's conscious intent to cause the damage. In other words, it wasn't enough under the CFAA to show that the defendant was technologically sophisticated and should have known that damage would be caused—the defendant had to want to cause damage.

The court decided that Resultly didn't intend to cause any damage to QVC's server, and therefore Resultly did not violate the CFAA. The court determined that QVC's crawl rate of up to 40,000 hits per minute was insufficient to show that Resultly intended to harm QVC's servers because Resultly's procedure had never caused a problem in the past and because QVC could have specified a slower crawl rate, but didn't do so.

The court emphasized that Resultly's business depended on the QVC website running smoothly as well as QVC allowing Resultly to crawl its site. Based on this evidence, the court concluded that Resultly couldn't have intended to damage QVC's website.

3. Facebook, Inc. v. Power Ventures, Inc.

Facebook, Inc. v. Power Ventures, Inc. was a lawsuit brought by Facebook alleging that Power Ventures collected user information from Facebook and displayed it on their own website. Facebook claimed violations of the CAN-SPAM Act, the CFAA, and the California Comprehensive Computer Data Access and Fraud Act ("CCCDAFA"). According to Facebook, Power Ventures made copies of Facebook's website during the process of extracting user information. Facebook argued that this process causes both direct and indirect copyright infringement. In a counter-claim, Power Ventures alleged that Facebook engaged in monopolistic and anti-competitive behavior by placing restraints on their ability to manipulate users' Facebook data, even when those users' consent was given.

The appellate court held that Power Ventures violated the CFAA, but only by continuing to access Facebook without permission *after* receiving Facebook's cease and desist letter. The court stated that violation of a website's terms of use or other computer use restrictions, without more, cannot give rise to liability under either prong of the CFAA's "access" prohibition. The court held that CFAA liability exists when a defendant has no permission to access a computer or when such permission has been revoked explicitly, and neither technological circumvention measures (such as like switching IP addresses) nor using a third party to access the computer can change this result.

Throughout its opinion, the court gave little consideration to Facebook's terms of use, and by extension, to Facebook's reliance on a terms of use to limit entry to an otherwise accessible site. Quoting **United States v. Nosal**, the court noted that, "not only are the terms of service vague and generally unknown…but website owners retain the right to change the terms at any time and without notice."

According to the court, Power Ventures initially had implied consent to access Facebook, since it had the consent of Facebook users. Thus, Power Ventures did not initially access Facebook without authorization. Facebook's cease and desist letter changed that, however. The court held that Facebook's mention of violations of the terms of use in its letter was not by itself sufficient to create CFAA liability. The court also noted that simply bypassing an IP address, without more, would not constitute unauthorized use of a website, since a user may not know he or she has been blocked or else the user may not associate an IP address block with a revocation of access directed at that specific user. After Facebook issued the cease and desist letter, permission from the Facebook users alone was no longer sufficient to constitute authorization to access Facebook for purposes of the CFAA.

The court reasoned that the scraping of a webpage inherently involves the copying of that webpage into a computer's memory in order to extract the underlying information contained therein. Even though this "copying" is momentary, it is enough to constitute a "copy" under copyright law and therefore infringement.

In summary, the court found that scraping data from Facebook profiles with consent from users, but not Facebook itself, constituted a CFAA violation.

4. hiQ Labs, Inc. v. LinkedIn Corporation

hiQ Labs' ("hiQ") business involved scraping data from public LinkedIn profiles and running an algorithm to determine the likelihood that specific employees may stay with the their current employer or may be seeking other employment. This data was then sold to the employers.

Although hiQ had been performing this scraping for several years with LinkedIn's knowledge, LinkedIn had a change of heart and issued a cease and desist letter, threatening legal action and blocking hiQ's access. hiQ responded by seeking an injunction. LinkedIn claimed that continued scraping was a breach of its User Agreement, and a violation of the CFAA among other laws. Regarding the CFAA claim, the court importantly stated that: "A user does not "access" a computer "without authorization" by using bots, even in the face of technical countermeasures, when the data it accessed is otherwise open to the public."

In this case, (1) hiQ did not have to login to LinkedIn in order to scrape the data because the profiles scraped were publicly available, (2) the profiles contained user-generated content which was owned by the users, not LinkedIn, and (3) LinkedIn had previously permitted hiQ's activities and its blocking hiQ might drive hiQ out of business.

If the data could only be scraped after logging in with a password, which it was not, the court held, then there is likelihood that the court would have found a CFAA violation. The court also held that that the application of technical blocking measures would not result in a conclusion that a user implementing countermeasures to continue access would constitute unauthorized access under the CFAA.

An additional important factor in the court's granting a preliminary injunction was that LinkedIn was the sole source of data for hiQ and allowing LinkedIn to block hiQ's access would put hiQ out of business and result in irreparable harm. The court ruled in favor of hiQ, and LinkedIn's appeal is currently being litigated.

Summary

The legality of data scraping requires examining: (1) whether the terms of service prohibit data scraping (and whether the terms of service are binding on users), (2) whether a password is required to access the website data, (3) whether copyrightable content is being scraped (and whether the fair use exception applies), and (4) whether the data scraping causes damage to the website.

Entities contemplating using scraping programs to access a public web site should also consider whether such action is authorized by reviewing the terms of use and other terms or notices posted on or made available through the site. A website seeking to prohibit data scraping should clearly prohibit data scraping in its terms of service and access to the data should be password protected.

Recommendations

In light of the above discussion and review of the recent developments in US courts' view of the penal and civil legality of web scraping, below our some general guidelines which should mitigate exposure.

1. No passwords or barriers should be broken to retrieve content being scraped.
2. The content being scraped should not be copyright protected.
3. If copyright protected, the scraped content should adhere to fair use standards.
4. The act of scraping should not burden the services of the website being scraped.
5. The scraper should not violate the Terms of Use of the site being scraped.
6. The scraper should not gather sensitive user information.

This document provides a general summary and is for information purposes only. It is not intended to be comprehensive nor does it constitute legal advice. If you are interested in obtaining further information please contact our office at:

Schuman & Co. Law Offices
Beit Bynet, 8 Hamarpe Street, P.O.Box 45392
Har Hotzvim, Jerusalem 97774 Israel
Tel: +972-2-581-3760, Fax: +972-2-581-5432
Schuman@schumanlaw.co.il
http://www.schumanlaw.co.il/