**Apartments.com Web Scraper Implications**

**Kyle Kofsky and Spencer Kaufman**

This paper is intended to act as a complementary resource to the related web-scraper tool and data repository. Firstly, it will define scrapers and bots in various contexts and provide a brief and overview of the implications of web-scraper tools and bots. This includes possible threats and vulnerabilities, as well as reasons for websites to take a defensive stance against bots and scrapers. Finally, we will discuss our reasoning for using Apartments.com as a basis, along with our conclusions and suggestions

Before discussing further details, we believe it is important to define what a web-scraper is. Webscraping is the act of data mining from websites by using software that simulates human behavior to extract information and data. In turn, this information allows for data to be warehoused before being indexed and analyzed [1]. From this, a vast variation of conclusions can be drawn, including behavioral, quantitative, and qualitative factors. Of course, the type of conclusions and information collected depend on the purpose of the data collection initiative. These can range from malicious to beneficial, while others may simply seek to gain a competitive advantage over a competitor.

An example of a “beneficial” web crawler or scraper is the Googlebot. This is a generic term for Google’s web crawler which comes in mobile and desktop variations [3]. Both versions simulate user behavior “to automatically discover and scan websites by following links from one webpage to another” [4]. We highly recommend following reference 4 to further investigate Google specific terminology and uses, while providing similar context to our web-scraper tool. After Google collects all of this information it builds a searchable index that essentially acts as the framework for the Google search engine that we know today[3].

Malicious web scraping uses have been consistently increasing in popularity and have more than doubled in recent years [2]. An example of a malicious use of webscraping is to search for discounts on online storefronts. These tools can be used to repeatedly scan for certain items across vendors and create a notification once a price drop occurs. While a human could perform this action, a bot can perform the same action on a much broader scale. This tactic is commonly utilized in gray-market sales, as an individual may use a bot to find a price drop and completely deplete the inventory from the vendor [1]. In turn, they can resell these products for market value and ultimately harm the vendors business. One common example of this surrounds concert ticket sales, which are often resold for several times the original price. While this occurrence is not business-ending, repeated use and abuse certainly has negative impact on e-commerce regarding reputation and customer flow. The line between this malicious use and gaining a competitive advantage is more nuanced than one may think. In fact, the process is almost identical but generally omits re-selling. Companies commonly use web scraping tools to quickly discover promotions and campaigns of their competitors and react more efficiently than they otherwise could. While this is not inherently malicious, it certainly makes the marketing and strategy field much higher paced and agile than it has ever been.

When designing a web scraping tool, we kept the aforementioned use cases and ideologies in mind to determine what website may provide useful information. Our initial thought was to attempt to scrape data from popular footwear and clothing retailers. However, popular digital storefronts such as these have detection mechanisms in place to prevent bot use. Consequently, we continue to search for vulnerable websites and settled on Apartments.com. In our initial design, we concluded that the most useful information would be drawn from the price, number of bedrooms, and zip code based on a user search input. However, an error in the naming schema for number of bedrooms did not allow us to collect this information. The final product was a scraper capable of collecting the name and full address of the unit, along with the price and data collected. This could then be analyzed to draw conclusions on unit pricing by region, and perhaps even infer bedroom count (despite the omission of this variable).

There are several key implications of this web scraper successfully collecting a vast amount of data from Apartments.com. The first of which being the fact that an anti-bot security measure is either not in place or is poorly configured. While this website simply acts as a means of listing property prices and not as a storefront, this leaves plenty of valuable data vulnerable to abuse. This leads into our second conclusion, that the use of a scraper allows for functionality that is likely intentionally excluded from the website. One such function is a listing of property and unit price history. If a user collected this data over several months or years, they could draw clear conclusions on which properties tend to “price gouge” and further capitalize on unsuspecting consumers. This could lead to a tremendous amount of negative publicity for property management companies, which may ultimately harm the reputation of Apartments.com. In a worst-case scenario, a malicious actor could also collect affiliated contact information from the website. This actor could then spam a property with emails whenever a price drop occurs to potentially reserve or draw false interest in pricing schemes. This could then result in inflated price increases due to falsely inflated consumer engagement metrics. This user would gain little to nothing from such action aside from harming others, unless they are an unethical competitor seeking to undercut their competition.

While none of these conclusions are specifically data driven, they act as a proof-of-concept showcasing the creativity of malicious users. The lack of bot prevention on Apartments.com, a high-profile listing site, is almost certainly not an accidental oversight. It is possible that they do not see ways in which their data could be used maliciously against them. However, this data is certainly open to abuse that would directly impact properties that use the website. If it is made public that a foundational prevention was omitted from their website, Apartments.com would certainly see a downturn in user traffic and reputation.

**REFERENCES**

1. <https://blog.radware.com/security/2016/03/good-bad-and-ugly-web-scraping/>
2. <https://www.infosecurity-magazine.com/news/data-theft-watch-web-scraping/>
3. <https://support.google.com/webmasters/answer/182072?hl=en>
4. <https://support.google.com/webmasters/answer/1061943?hl=en&ref_topic=9426101>