**ABSTRACT**

The main goal of our research was to find a way to access Dark-Web for scraping data. Because the Dark Web has generated great interest from academics and governments who have sought to unveil the identities of participants in these highly lucrative, yet illegal, marketplaces. Traditional Web scraping methodologies and investigative techniques have proven to be inept at unmasking these marketplace participants. For this reason we tried scraping the Dark Web Data with free tools found on the World Wide Web. Our main obstacle throughout this research was we could not work on Linux operating system as have researched that Scraping Dark- Web is more flexible on Linus rather than other OS. Though we tried on Windows and Mac with different methods. In this paper we will discuss about the reasons for why we were not able to scrap data.

**INTRODUCTION**

Surface web or World Wide Web is relatively easy to traverse, using traditional Web browsers, like Microsoft’s Internet Explorer or Edge, Google’s Chrome or Apple’s Safari. However, the World Wide Web only accounts for a portion of Internet trafﬁc, while content on the Internet that is unindexed is referred to as the Deep Web. The unindexed portion of the Internet that is intentionally hidden and inaccessible by standard Web browsers is referred to as the Dark Web. The Dark Web is an encrypted network that utilizes the public Internet. Among the various overlay networks, such as Garlic Routing or Tunnel Routing, the most widely used one is Tor Routing, which was originally developed as part of a secure communication effort of the U.S. Naval Research Laboratory, to protect and anonymize trafﬁc by passing it through multiple layers of encrypted. The Dark Web is purposely hidden using a peer-to-peer (P2P) network and Dark Websites are primarily accessed using the Tor Browser, which is a user-friendly browser that protects the anonymity of the user, and can be important for individuals seeking to overcome censorship, ensure their privacy, or for criminals who seek to obfuscate their identity. According to the Tor Project’s Website, “Tor is free software and an open network that helps you defend against trafﬁc analysis, a form of network surveillance that threatens personal freedom and privacy, conﬁdential business activities and relationships, and state security”.

**The Surface Web Marketplace**

Scraping data off the surface web marketplace was our very first task of these research and we had a few hiccups during this part. Our first step was surfing through these marketplaces to do a have a conceptual idea of how these sites work. In this process we started by going through a few sites like eBay, Amazon, AliExpress etc. While going through these sites, we were quite surprised by how these sites actually differ from one another. We categorized these sites on a few categories. The categories being:

**1. Official Product Sellers:** Sites like Amazon or Newegg mostly sell official products. These pages are far more organized. They have the more pristine look with users not having to go through a lot of searching for getting to the page they want to go to. And as each product is official, the hassle of finding the best deal for the same product is gone. But at the same time, having to place the order with only a single choice per product can drive people away. So, Amazon does provide in some cases the choice of choosing sellers as well. The users can choose which seller they want to buy a product from. But the items with more sellers are seldom found. But no matter what, these pages do not sell opened products. They ensure that the products are untouched and not tampered with in any way. In that prospect, there is no scope of any scams.

**2. Official/Second-hand Product Sellers: sites** Websites like eBay and AliExpress were more user oriented and allowed for there to be second hand items as well as official branded products. There were differences between these second hand stores as well. For example, AliExpress allows users to pick “colors” which means for the user to be able to sell different products in the same listing while eBay has pages for each of the colors. Which means eBay is a bit more structured than sites like AliExpress. Sites like them do have more things on common though. Like the process of finding the best deal for the product you are trying to buy. Which could be seen in both ways. In the first hand, having more options is always good. But on the other hand, too many options can make it quite frustrating for the customer to find what they need. There is a big chance of customers to find better deals after already placing an order which can always be a very disheartening discovery. There are also some chances of scams that could turn us off these sites.

**3. Branded Shops:** These sites work as dedicated pages for specific brands. This allows the brands themselves to cut out the middle man to sell their products directly. This also removes however little confusion that could be left on sites on the 1st category. It also gives the customers a more clear-cut catalog of their products. Only downside of these sites is customer traffic. As more buyers are likely to flock to a dedicated selling website like Amazon, not many visit the branded sites. Also, as these sites are brand specific, it strips away the choice of brands from the customers, who are more inclined to look through different brands to fin the one that suits them personally.

**4. Subscription Based:** There are some sites like Loot Crate, Doller Shave Club and many software selling ones that sell these items in a subscription basis. Instead of a static cut, these sites periodically charge buyers to provide them with certain goods or service. These types of services seem like a new way of selling products that have gotten quite popular in the last 5 few years or so. These have the same problems as branded shops to, only to a higher degree. Most of these sites provide a singular service or at most a binary amount of options for the customers which though seems bad, have proven to be quite successful over the years.

**The Deep Web Marketplace**

The marketplace of the dark web that we were able to surf through was quite peculiar to say the least. To categorize these sites is a hard task as the ones we managed to visit were strikingly similar. Most of these sites were selling illegal things like drugs, ammunition, weaponry, counterfeit bills etc. The sites had unintelligible links that appear to stop working every now and then. As a form of countermeasure perhaps, these sites seldom uses the same links. Changing every once in a while to keep the anonymity. But the more interesting part was, the structure of these sites was very much the same. These sites are very one dimensional. Working as branded websites, each of these sites were shoddily created to usually point to a singular seller of very specific products. For example, there were a number of drug sellers that all essentially sold the same products but each of these sites had only one page with no more than a decimal number of products. Same case with the weapon and counterfeit bill selling sites as well. We started seeing a pattern with the same template of a sloppy background with a few products in a singular page being on almost every single page. It almost felt like all of the pages were created by a single person, or at least follows the same template. As we later discuss, we failed to scrape these sites. But to investigate further we tried downloading whole pages as html files and after seeing their html structures, we were more curious about that factor. We also saved the sites as text to verify the patterns without having to get into the dark web every time. We attached a few of these saved files with the research. It’s fascinating to see so many of the deep web marketplace use the same template for their pages. But again, the scope of our research could have been small too small to cover it. We used sites from The Hidden Wiki and another hidden wiki page that had a few more sites than the other. A very important discovery about the deep web marketplace was the use of bitcoins. Every single marketplace we visited had bitcoins as its only currency of choice. This improves anonymity as it puts all of the buyers and sellers in the same group of bitcoin owners instead of location-based currencies.

**Methodology**

**Data Scraping**

A big part of our research was supposed to be the scraping of both the Surface and Dark Web, which later became the most frustrating part of our research as well. We knew the conventional methods of scraping was to use Python plugins. Scraping the surface web was not the frustrating part as we had help from many tutorials and documentations. We used the python plugins Beautiful Soup and Scrapy to varying degrees of success. What we found out during scraping these sites was quite surprising. The html files of the major surface web marketplace are quite disorganized. Sites like eBay, Amazon, AliExpress have conflicting headers and id names that makes scraping these sites harder than other sites. We could not confirm if it was a planned effort to limit the scraping capabilities, but it did hinder us to the point we had to find other marketplace sites to scrap. We chose to scrap the computer parts site Newegg. Newegg was a lot easier to scrap as it had clear headings for product listings. We managed to scrap a single page by using Beautiful Soup. As well as scraping a site using Scrapy. The codes of which are listed down below.

And with these miniscule bits of successes to our belts, we focused our attention on what was arguably the most challenging part of our research, scraping the dark web.

**Scraping the Dark Web**

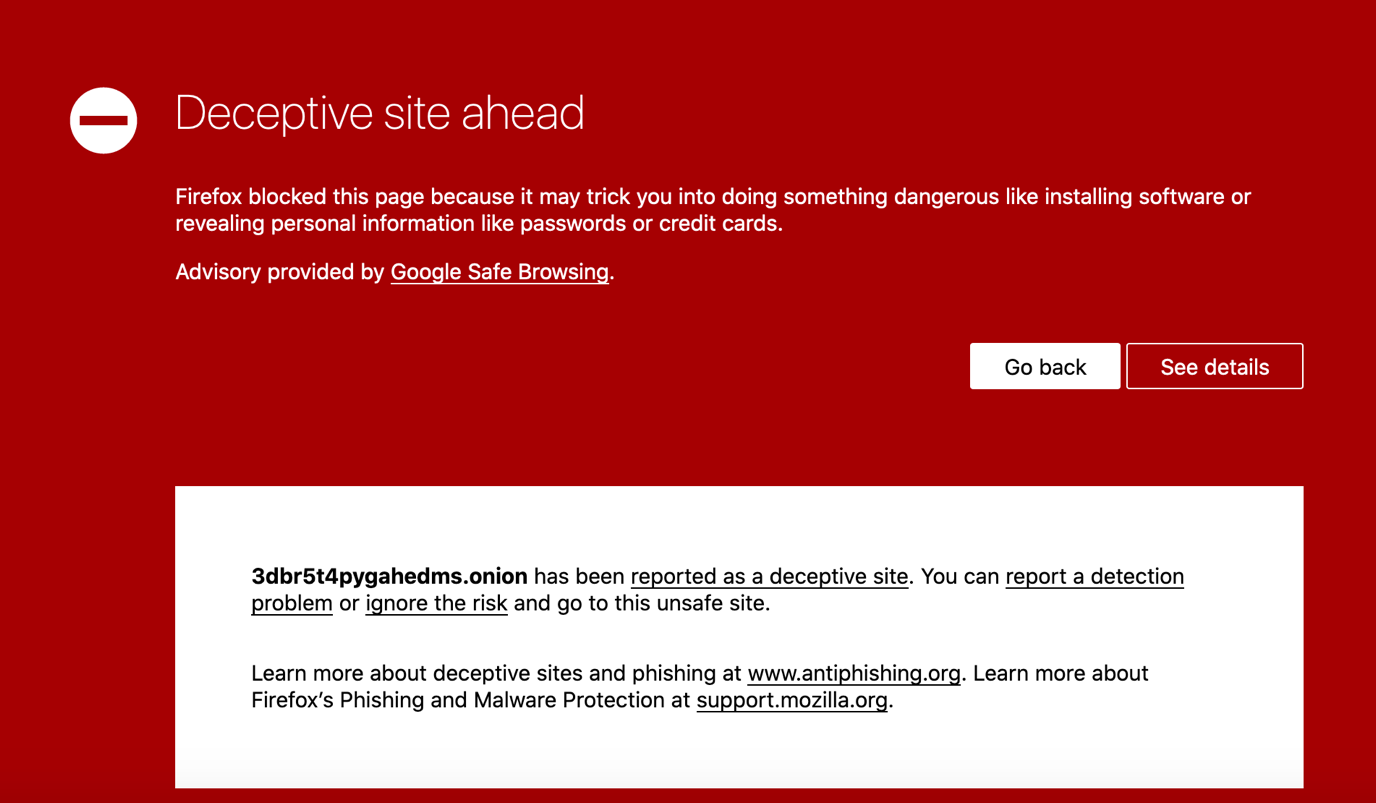
Scraping the dark web was undoubtedly the hardest part of our research. Especially as we had no prior knowledge of scraping. But with the small amounts of testing we did on the surface web we thought it was going to be a doable task. But to get into the deep web, first we had to get into the deep web.

**Getting into the deep web:** To enter the dark web, the process was actually a lot easier than we were led to believe. We used multiple methods to accomplish this task with most of them being quite successful. For our first method, we tried the most conventional of ways. We installed a free VPN software for our PC. Getting a free VPN software is as easy as searching “A free VPN software” on Google. Software like Tunnelbear, Windshare as well as trial versions of NordVPN were the ones we tested which all gave good results. But a paid option is definitely recommended for maximum security. We used any one of these VPN software to bounce our IP address. We then downloaded the TOR browser. It is also available for download on the internet. But, it is recommended to only download the one that comes up first in the searches as other versions can be scams or viruses. After turning on the VPN software, all we needed were dark web links for the TOR browser. We found these links at The Hidden Wiki. The site provided us with quite the assortment of dark web links. We decided to do our research on links found on this site. We tried BeautifulSoup as our method of scraping.

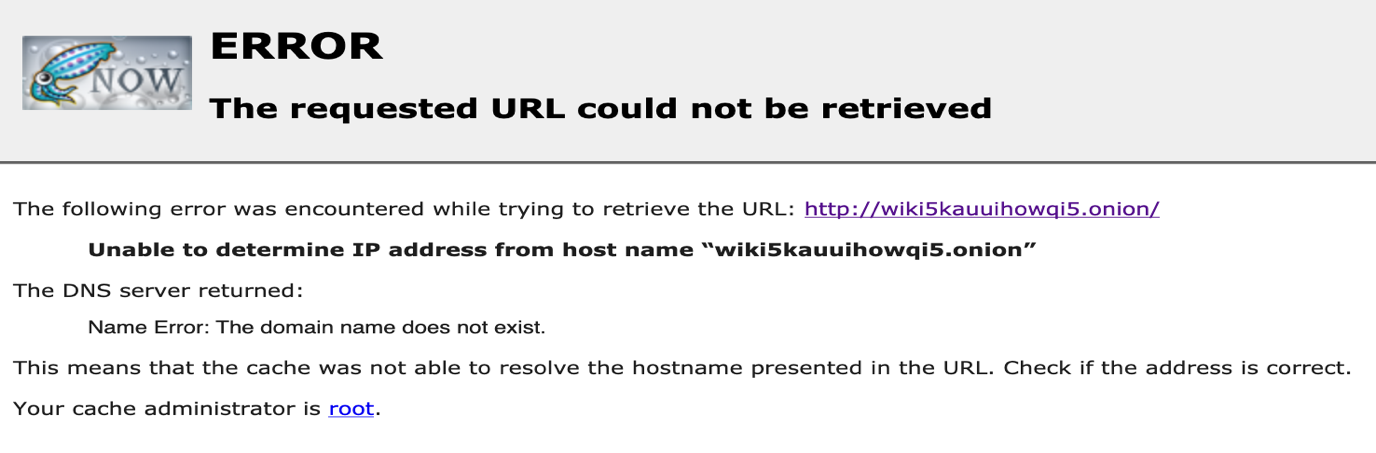
**Scraping with Beautiful Soup:** Our frustrations began with trying to scrape the dark web with this method as this was ultimately a fruitless attempt. We began by trying to scrape it similarly we scraped Newegg. But we fell at the very first hurdle as the URL of the dark web sites were not supported with the URLOPEN method that we found was at the basis of most Beautiful Soup scrapings. As that wasn’t working, we tried out a few different methods like using python to set up TOR as a targeted browser. But we couldn’t get that to work either. This was the end of our efforts with Beautiful Soup and our first idea of how scraping the dark web was going to be like.

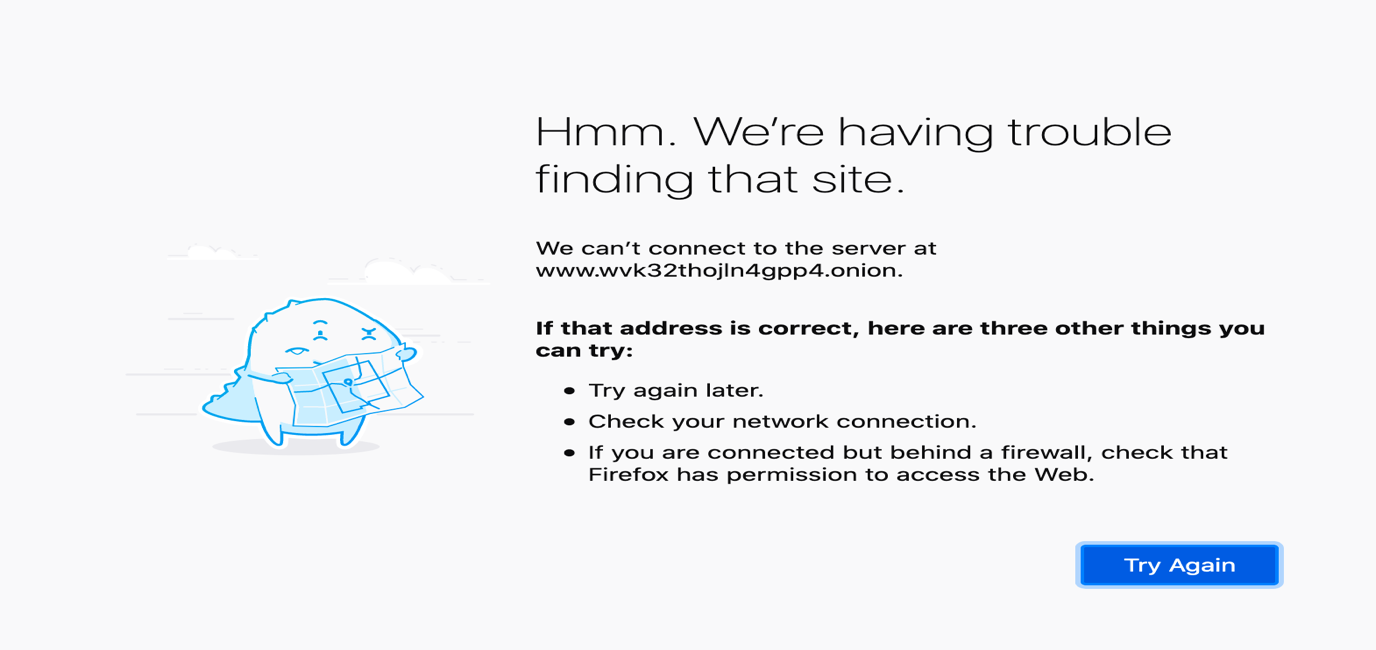
**Scraping with Scrapy:** This was the method that showed us that python was not the best for scraping the dark web. Python is best used to do work with http or https URLs. We could manipulate it a bit but it would still be fruitless. For example, the “URLOPEN” tool is a freely available command line utility that is widely used for http and https browsing. But we could use tool like cURL, to build a Tor wrapper around that command line tool, which can theoretically allow cURL to take advantage of the anonymization of Tor. However, even if we were successful in doing so, this process would still only provide (anonymized) access to the World Wide Web. The Dark Web .onion sites, which do not implement the standard Domain Naming System, would still be out of reach. With this devastating realization, we did not try further with scrapy as the main part was getting into the URL itself which we could not achieve in any capacity using python at all.

**Scraping with Selenium:** By changing Firefox configuration we were not able to scrape the dark web either. Typing about:config on the search bar and locating network.dns.blockDotOnion setting and changing its value to false, we were able to access dark web before from firefox with vpn turned on. Some .onion sites were blocked by firefox as users reported them as deceptive sites. The following message was shown by firefox.



In case of other .onion sites, simply firefox couldn't connect to their server or retrieve their url. Firefox showed following messages.





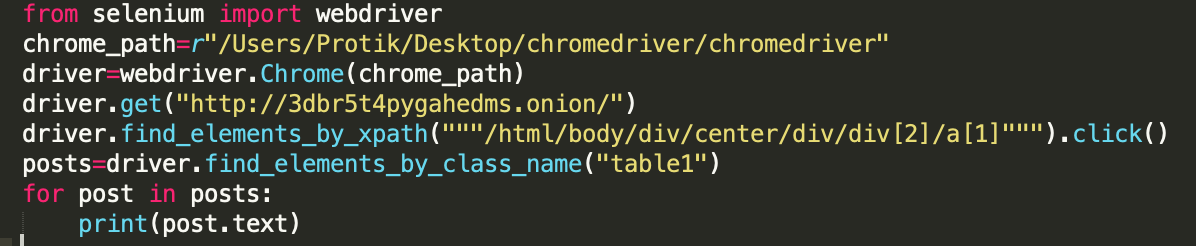
So, we were not able to scrape any webpage via firefox. Then, we moved on to attempting to scrape dark webpage via chrome.

**Attempt to scrape using selenium webdriver:** We attempted to scrape dark webpage using selenium. We planned to access the dark web over tor network. Selenium webdriver can't work on tor browser so we used chrome web browser. To access tor network via chrome browser we used a chrome extension named Tor Control (Anonymity Layer). This extension allows to access tor network via other web browsers like chrome, firefox etc. To begin with, we installed a vpn software and selenium through mac terminal. On mac terminal, the command to install selenium was-

****

Then we installed the chrome extension from the following link- <https://chrome.google.com/webstore/detail/tor-control-anonymity-lay/kjoabfljeghcinlpjhdbdfbcflapkccm?hl=en-GB>

After that, we ran the VPN software, enabled the extension in the browser and tried running lines of python code to scrape data from a table of dark webpage. The following python code was executed-



We couldn't scrape the webpage because the .onion site couldn't be successfully reached. After running first 4 lines of code, browser opened a new window automatically and tried to load the dark webpage but after a while "This site can't be reached" message was shown in the chrome browser. We tried other sites but the responses were the same. From this we assumed that most probably the extension could not connect to the tor network properly. So the browser could not load .onion sites.

**Results**

In the end, we were trying our best to try and scrape the dark web marketplace so much that we could not focus on the main task. We still wanted to make a fair comparison between the surface and dark web marketplaces. But without the scraped data, our options became a lot more limited. So, we then resorted to manually saving the contents off of the dark web. We believe the data we saved can be enough to make a fair comparison between the two marketplaces. A few of the comparisons we came up with in the end are:

* The surface web marketplaces have a lot of different designs which can be divided into many categories whilst the dark web marketplace mostly has the same layout.
* A few of the surface web marketplaces have atrociously unorganized html files. On the other hand, the dark web marketplaces usually have quite a primitive layout which makes it quite organized.
* Surface web marketplaces have confusing html files that makes it hard to scrape the html files themselves. But the dark web sites have encrypted domains that makes it hard to connect with the URL itself. Though it’s trivial html files make it seem very much possible to scrape it with any of the scraping tools.
* The surface web marketplace (if not country based) has choices when it comes to currency. While the dark web only operates on crypto currency (bit coin).
* For Surface Web, one can know and contact with the seller or buyer apart from internet but in dark web, the both party (Seller & buyer) remain undercover or anonymous.
* The Surface Web’s sites stay for a long time but the dark web’s site frequently changes it addresses.

