Data Capture and Analysis of Darknet Markets

Australian National University Cybercrime Observatory

Matthew Ball, Roderic Broadhurst[1], Alexander Niven, and Harshit Trivedi

March 2019

Abstract

Darknet markets have been studied to varying degrees of success for several years (since the original Silk Road was launched in 2011), but many obstacles are involved which prevent a complete and systematic survey. The Australian National University's Cybercrime Observatory has developed tools to collect and analyse data captured from the darknet (illicit cryptomarkets). This report describes, at the high level, a method for collecting, and analysing, data from specific darknet marketplaces. Examples of typical results that may be obtained from darknet markets and current limitations to the automation of data capture are breifly outlined. While the proposed solution is not error-free, it is a significant step in the direction of providing a comprehensive solution tailored for data scientists, social scientists, and anyone interested in analysing trends from darknet markets.

Introduction

Illicit cryptomarkets (or darknet markets) are e-commerce style websites specializing in the sale and distribution of illicit content. Typical products offered on darknet markets include: drugs, pharmaceuticals, identity documents, malware and exploit kits, counterfeit goods, weapons, and other contraband. Europol estimates that two-thirds of the products listed on darknet markets between 2011-2015 were drug related and these markets are '..one of the engines of organised crime' in Europe (EMCDDA & Europol 2017: 15; Ball, Broadhurst & Trivedi, 2019).

Darknet markets are only accessible through tools such as TOR ("The Onion Routing" project). TOR specialises in websites ending in the '.onion' suffix, which does not use the standardised Domain Name System (DNS) for website retrieval. This bespoke system utilizes a blocker point between the website and user, ensuring neither knows the identity of the other. As a result, the identities and locations of darknet market users remain (pseudo)anonymous; they cannot be (easily) tracked due to the layered encryption systems used (Holt & Bossler, 2015). Although it is not the original motivation for these network tools[2], a substantial share of the Internet traffic operating in the encrypted 'deep web' is comprised of darknet markets.

Darknet marketplaces mimic clearnet e-commerce services such as eBay, with three main actors: vendors, buyers, and market administrators. Buyers can leave reviews, message vendors, and dispute transactions. Vendors provide product descriptions and basic details such as quantities, prices, and shipping services. Administrators typically receive 3–8% commission from each sale and provide escrow services and overall supervision of the website and market operation.

Aim and Scope

The aim of this report is to provide a reliable and reproducible method for scraping darknet marketplaces for relevant products and users who visit such sites. Within a Darknet marketplace, the product and vendors are topics of interest in research and analysis. Description of the buyers of these products is limited, partly due to the scarcity of data about buyers. The data captured from this method can then be analysed to discover trends of many aspects of the market including purchases, users, and the supply and demand of products. The Australian National University's (ANU) Cybercrime Observatory proposes a method for 'crawling' (the automated act of exploring a site to discover and download website pages) a marketplace, and 'scraping' (the automated act of analysing a HTML file to export information) relevant information for analysis. While the methodology provided is illustrated on the Apollon Darknet market, the method can be generalised to other marketplaces. Due to the different nature of addresses of the product images present, there is no ready method of acquiring the product images on each of these websites. Their direct .onion link/s are instead saved for later download.

For the purpose of this report, we have trialled the proposed method on the Apollon Darknet market. The ANU's Cybercrime Observatory uses this method for any Darknet marketplace data capture in the TOR environment. In the Observatory's research, a market is considered omnibus if there are over 1000 product listings within the history of the marketplace. General marketplaces are considered priority over product specific marketplaces and are the markets currently explored.

---

[2] The primary motivation for the development of TOR was to protect government communications as described on the project homepage: https://www.torproject.org/about/torusers.html

The Apollon Darknet market is a small market on the dark-web, easily accessible for research purposes. Therefore, a proof of concept using the Apollon market is provided below. An appendix is included which details the Apollon market.

Previous Work

While work on collecting and analysing darknet market data is an evolving subject area, there has been progress in reliable data collection methods for researchers. Décary-Hétu, D., & Aldridge (2015) and others have proposed the 'digital traces' of actual drug selling data collection method as demonstrated by many papers in 2016 that described darknet markets by collecting data from Tor sources (e.g. Décary-Hétu & Aldridge 2016; Décary-Hétu, Paquet-Clouston & Aldridge 2016), representing the methodological evolution from traditional administrative data to so-called 'full population' or 'webometric' approaches (Barratt, & Aldridge, 2016, p 3; see especially volume 35 of the special edition of the International Journal of Drug Policy, 2016). The benefit Barratt and Aldridge (2016) note "…is the ability of researchers to generate knowledge based on populations – of drug sellers, of drugs sold, of drug prices of drug quantities" rather than on potentially unrepresentative samples of arrested offenders or drug seizures. Application of this methodology does require interaction with illicit marketplaces and merits ethical consideration of the implications and risks for darknet users, sellers and others, including researchers (Barratt & Aldridge, 2016)[3]. As such, a transparent approach has been adopted to promote replication and validity of the results presented. We believe this is vital to reduce potential inconsistencies in results by standardising the methodology used within research. Inconsistencies can occur through the instability of the TOR network, and the marketplaces in question (Van Buskirk, et. al., 2015) and as such, standardisation is vital. While an automated system such as this is not perfect, it is a significant improvement over the manual methodology previously employed (Dalins, Wilson & Carman, 2018).

In the wider subject area of general data capture on darknet websites, there has been growing concern for the 'human' aspect as researchers are required to interact with illicit items. The scale of data from these websites has increased significantly as popularity of the darknet increases, resulting in a requirement for automation and classification methods for data analysis purposes. While many classification models exist, specialised models have now become important as an absence of a taxonomy capable of recording such data can be detrimental to end users of the information required (Dalins, Wilson & Carman, 2018). Adding to this, illegal information such as child exploitation materials (CEM) can be prevalent on darknet websites and the development of automated classification methods can significantly reduce the amount of images viewed by individuals who classify and investigate CEM (Dalins, Tyshetskiy, Wilson, Carman & Boudry, 2018). The health impacts of viewing CEM specifically are well documented (Burns, Morley, Bradshaw & Domene, 2008) and a reduction in interaction with CEM is a significant advancement in the field.

Method

Darknet sites are usually not crawled by generic crawlers. Partly because the web servers are hidden in the TOR network, and partly because they require the use of specific protocols to be accessed. Although, in the generic case, crawlers (and scrapers) are typically

---

[3] ANU Ethics Approval 2019/45, "Policy Sprint: Current trends in fentanyl availability on crypto-markets", was approved for the development of this research methodology. The *National Statement on Ethical Conduct in Human Research (2007)* (National Statement (2007) are guidelines for researchers made in accordance with the *National Health and Medical Research Council Act 1992*: available from: https://nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018

designed and implemented in a single script, the Observatory's team has found it more convenient to keep the implementation and control of both the crawlers and the scrapers separate. The reasoning for this decision is described in Section 5 Limitations. Accordingly, the processes for the crawlers and scrapers, as previously defined, have been described separately below.

Crawler

A crawler is an automated script defined to search the entire site in a methodical manner, and find as many unique pages as possible to create a static copy of the marketplace website for later analysis. Static copies of websites produce a timestamp for the time of capture. The static copy retains the structure of the link and allows users to navigate the static site as if browsing in real time (Dolliver and Kenney, 2016).

The ANU's Cybercrime Observatory has written a specialist set of crawlers designed to capture product and vendor information from a collection of darknet markets. The process through which these crawlers operate is typically a three step process: Perform a "page crawl" of a listing of products (and scraping of the page(s) for later purposes); perform a "product crawl" based on the scraped listing pages; and finally, perform a "vendor crawl" based on the scraped listing pages. This design depends upon necessary page information, which is not given in all cases. For the few cases in which this information is not available, a different approach is taken. In some specific cases, products are denoted by their product listing number and as such eliminate the need for a initial page crawl. This implementation is market specific and does not hold in all cases.

This data capture process is limited by various factors, including DDoS protection of marketplaces. These limitations are further described in the section on Limitations and Future Work.

Scraper

HTML scraping is a technique for extracting large amounts of data from websites. This data is exported to a CSV file (comma-separated value file) for later analysis.

The previously defined crawlers have a corresponding scraper to collate information from HTML files and export the information to a human readable format. This also acts as a barrier to filter out incomplete or incorrect information. As with the crawler design, the three scrapers exist for two main topics of information: the product in question, and the vendor. In line with the crawler, there are cases where variations in the presentation of listings can eliminate the need for an initial page scrape.

Although scrapers for market sites are implemented differently, a subset of common features appear in each market, which is recorded in the corresponding CSV files. Each page scraper writes the Product Name, Vendor Name, Price in BTC, Price in AUD (calculated from the value of BTC at the time of capture), Product URL and the Vendor URL to the CSV file. Each product scraper writes the Product Name, the Vendor Name and the Price(s) to the CSV file. Lastly, each vendor scraper writes the Vendor Name and Vendor Rating to the CSV file. Other information such as quantity sold or a vendor's compliant index can sometimes be present on specific markets and scrapers can be adjusted accordingly to include these data points.

As with the crawler, the scraping process can be impacted by various factors including incomplete data capture or non-standardised forms. These are noted in the next section describing limitations and future research.

Results

For data analysis purposes, the Jupyter Notebook[4] application is employed. The Jupyter Notebook reads the data from a CSV file into a "DataFrame" (a two dimensional data structure; a column and a row of equal length). The CSV file can be used by many statistical packages. Typical analysis that the data enables is provided in the following examples. The first two examples illustrate a trend analysis of the available products and the number of vendors active over the trial period.

Jupyter allows a 'literate' style of programming which combines literate text as well as code. This allows the creation of documents which contain "the analysis description and the results (figures, tables, etc..) as well as executable documents which can be run to perform data analysis." (Jupyter Documentation, n.d). During this analysis process, no new data is added to the datasets. This process "wrangles[5]" the data into information which can be used to draw conclusions from the data. For instance, this method calculates the "unit price" for products listed with a weight (i.e. for any products listed in terms of grams/milligrams/etc.) and this allows the calculation of how much each item costs in a uniform weight scale. This also enables a "purity weight" to be applied (i.e. if an item is listed as 95% pure, the unit price is adjusted accordingly) and is illustrated in image 3.

Image 1: Example (Total Listings) Graph for Apollon Market between Dec 17 2018-Feb 25 2019. Note, the 1st Degree Trend and the 3rd Degree Trend lines describe the regression relationship between the time of scrape and number of listings modelled as a nth degree polynomial of time.

---
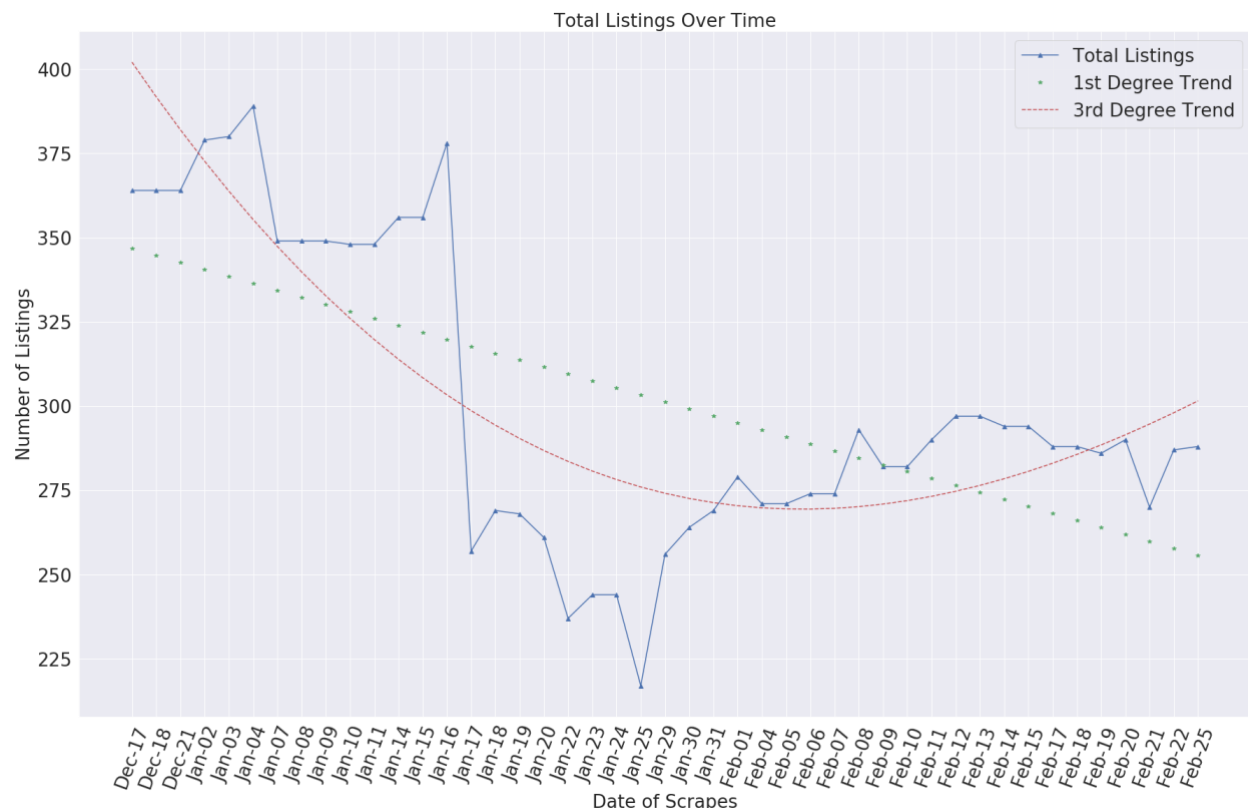
[4] The Jupyter Notebook is an open-source application that allows us to create and share live code-and-run notebooks, allowing for easy visualization and narratives. It supports over 40 programming languages such Julia, Python, R and Scala (Jupyter Documentation, n.d).

[5] Data Wrangling is the process of mapping and transforming data from its raw form to a more understandable form, that is ready for data analysis.

Image 2: Example Active Vendors Graph for Apollon Market between Dec 17 2018 - Feb 25 2019. Note, the 1st Degree Trend and the 4th Degree Trend lines describe the regression relationship between the time of scrape and number of vendors modelled as a nth degree polynomial of time. The green and red shadows along these lines describe a ±5% confidence interval.
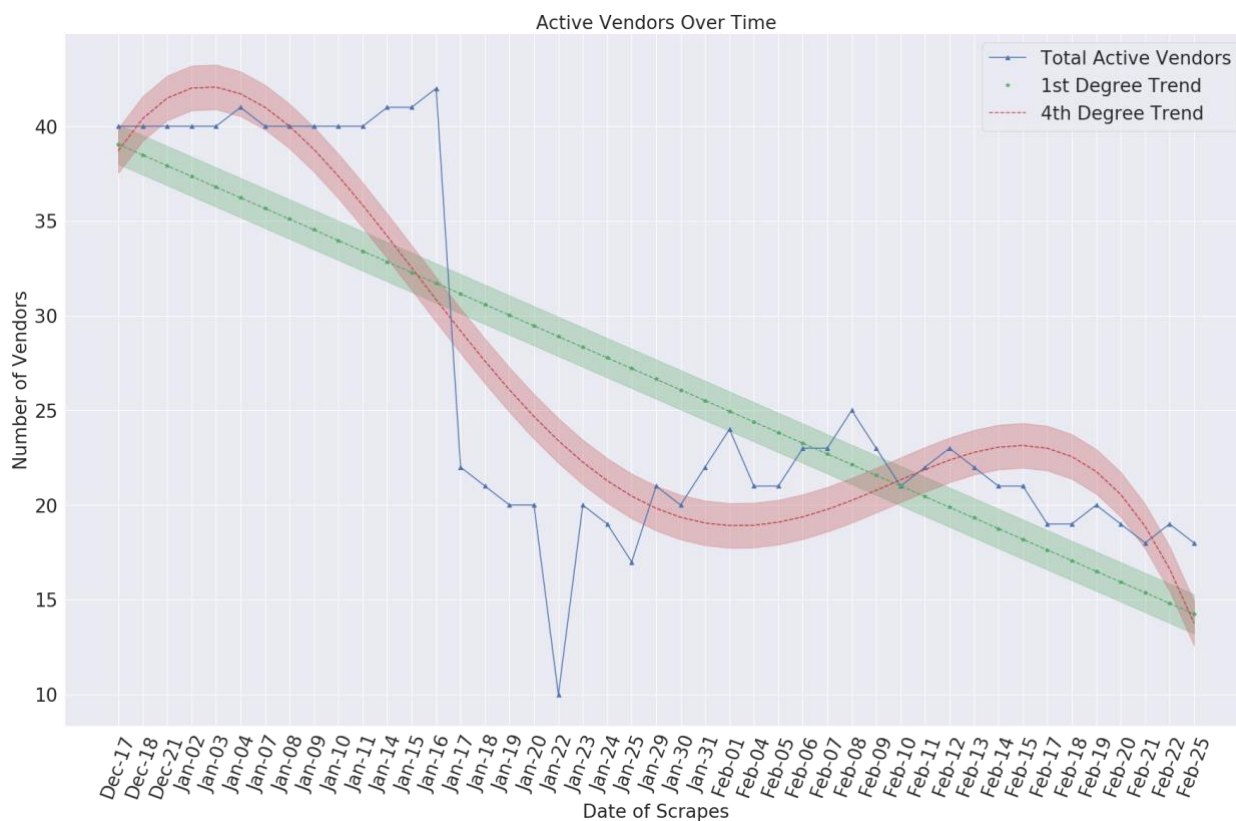
Image 3: Example Unit Pricing (ordered by Price in AUD, Units and Weights)

	Product	Vendor	Price in AUD	Weights	Units	Unit Price in AUD
0	★90%+ Pure Pink Speed Paste★ Qty: 1000g	HappyChemist	3257.04	0.90	1000.0	3.62
1	★90%+ Pure Pink Speed Paste★ Qty: 500g	HappyChemist	1850.59	0.90	500.0	4.11
2	★90%+ Pure Pink Speed Paste★ Qty: 250g	HappyChemist	1036.33	0.90	250.0	4.61
3	★90%+ Pure Pink Speed Paste★ Qty: 100g	HappyChemist	458.95	0.90	100.0	5.10
4	★90%+ Pure Pink Speed Paste★ Qty: 50g	HappyChemist	251.68	0.90	50.0	5.59
5	***(28g) 84% pure MDMA - $560 USD + MDMA Test ...	boogknight	11.91	0.84	28.0	0.51
6	20 gram amnesia haze	greenpirate	192.46	1.00	20.0	9.62

Limitations and Future Work

While the current iteration of darknet market tools are operational and usually reliable, there are several current limitations that will need to be mitigated.

Prolonged Downtime

Darknet Marketplaces are volatile and can at times experience significant downtime. There are multiple reasons that a darknet market may be down, including: distributed denial of service (DDOS) attacks, exit scams, voluntary closures, and occasionally hacks or de-anonymization or seizure by law enforcement (Van Buskirk, et. al., 2016). Recently, Dream Market experienced a downtime of over 11 days due to a rogue user DDoSing the service. This at times cannot be predicated ahead of time and can cause crawlers to fail or produce significant delays in a successful crawl of a marketplace. This issue requires researchers to constantly assess the execution of crawlers to ensure the programs are working as expected. Adding to this, two stage verification systems must be employed to redownload failed pages based on incomplete downloaded files.

Operational Security of Sites

In an effort to curb the potential of a DDoS event and eliminate the use of automated bots and other systems many darknet markets have employed the use of automatic timeout on accounts. As an account is required to view the product listings, the timeout can become problematic based on the time taken to perform a crawl. While this is not normally an issue on clearnet websites, the instability and speed limitations of TOR heavily impact the time consumption of market crawlers.

To mitigate this issue, the use of saved cookies is employed in an attempt to force the site into continuing to allow access. Scripts can also be developed to refresh the page in a TOR browser to show the cookie as active. The same process can also be performed manually, although is tedious as the number of marketplaces being crawled simultaneously increases.

Unique Cataloging of Information by Marketplaces

Due to the non-standardised implementation of marketplaces by developers, there are cases where data can be non-complete in a marketplace's listings. This can be caused by a variety of factors including: server-side issues where the marketplace does not hold the correct data, the crawler has not saved the data correctly, or the scraper cannot interact with the saved data in an expected manner. Markets also adapt to bans or the inclusion of other products, further causing issues with the crawler and scraper compatibility. This limitation cannot be predicated ahead of time and requires constant monitoring of scrapers to ensure data integrity is of an acceptable standard.

Currently, the scrapers are monitored and designed to report any faults for manual inspection and addition to the data set if necessary. Future development in this space could allow an automated error reporting and re-crawl system to be implemented.

Future Work

While the methodology proposed in this report provides researchers a reliable method to crawling and scraping darknet markets, there is further optimisation that can occur to the method. Currently, images are not downloaded due to the instability and speed of the TOR service, and the slowdown of services can impact on overall crawl times and possibly trigger the timeout function on many marketplaces. Further research could be done into isolating these links and separately downloading the images, as informal research indicates the images are held on a different service to the marketplace and could therefore could theoretically be reached without requiring login and DDoS protection workarounds.

On the subject of DDoS protection and their workarounds, there is significant advancements into the automated solving of 'CAPTCHAs' (Completely Automated Public Turing test to tell Computers and Humans Apart), with Google's reCAPTCHA becoming obsolete with advancements of Google's own speech and image recognition technologies (Sivakorn, Polakis, Keromytis 2016). Many darknet marketplaces use bespoke captcha systems, with relative success. While advancements have been made into trying to break these systems (Csuka, Gaastra & de Bruijn, 2018), generalisation across multiple captcha systems within the darknet would benefit crawling and scraping methods significantly.

Finally, if the captcha system employed cannot be broken, development of cookie refreshing would suffice as a workaround until an advancement is made. Research into a TOR or torify plugin using the same cookie to ensure the server still recognises the user as active would reduce the startup and maintenance time of the crawlers employed significantly.

Conclusion

While darknet marketplaces have been the subject of exploration for over 6 years now, there has yet to be a reliable method for capturing relevant data accepted as a subject area standard. This report proposes a methodology for capturing and analysing raw data obtained through the TOR network, specifically focussing on darknet markets. This foundation can be further built upon to generalise methodology for wider use within the TOR data capture space, or increasing stability of the existing frameworks.

The methodology described has been designed such that it can be utilized by parties with varying levels of technical knowledge, allowing greater access and inclusion to the data available. In its currently implemented form, drawbacks exist as constant maintenance is required due to the instability of such marketplaces, although this is an unavoidable feature of the subject area. This method has been used by the ANU Cybercrime Observatory on over 10 darknet marketplaces trialing successfully for over four months and can be adapted as necessary to niche and other sites on the darkweb. The methodology has been used to provide data inform government agencies within Australia and broader international community.

References

Australia, National Statement (2007) National Statement on Ethical Conduct in Human Research, available from https://nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018

Ball, M., Broadhurst, R. and Trivedi, H. (2019). Statistical Bulletin: How much fentanyl is available on the darknet?. Australian Institute of Criminology, Canberra.

Barratt, M. and Aldridge, J. (2016). Everything you always wanted to know about drug cryptomarkets* (*but were afraid to ask). International Journal of Drug Policy, 35, pp.1-6.

Burns, C., Morley, J., Bradshaw, R., & Domene, J. (2008). The emotional impact on and coping strategies employed by police teams investigating internet child exploitation. Traumatology, 14(2), 20-31. doi: 10.1177/1534765608319082

Dalins, J., Tyshetskiy, Y., Wilson, C., Carman, M., & Boudry, D. (2018). Laying foundations for effective machine learning in law enforcement. Majura – A labelling schema for child exploitation materials. Digital Investigation, 26, 40-54. doi: 10.1016/j.diin.2018.05.004

Dalins, J., Wilson, C., & Carman, M. (2018). Criminal motivation on the dark web: A categorisation model for law enforcement. Digital Investigation, 24, 62-71. doi: 10.1016/j.diin.2017.12.003

Décary-Hétu, D., & Aldridge, J. (2015). Sifting through the net: Monitoring of online offenders by researchers. European Review of Organised Crime. 2. pp.122-141.

Décary-Hétu, D., Paquet-Clouston, M., & Aldridge, J. (2016). Going international? Risk taking by cryptomarket drug vendors. International Journal Of Drug Policy, 35, pp.69-76. doi: 10.1016/j.drugpo.2016.06.003

Dolliver, D. (2015). A rejoinder to authors: Data collection on Tor. International Journal Of Drug Policy, 26(11), 1128-1129. doi: 10.1016/j.drugpo.2015.08.004

Dolliver, D. and Kenney, J. (2016). Characteristics of Drug Vendors on the Tor Network: A Cryptomarket Comparison. Victims & Offenders, 11(4), pp.600-620.

Csuka, K., Gaastra, D., & de Bruijn, Y. (2018). Breaking CAPTCHAs on the Dark Web.

Guide to Python (n.d.). HTML Scraping. Retrieved from https://docs.python-guide.org/scenarios/scrape/

Holt, T. J., & Bossler, A. M. (2015). The future of cybercrime, technology, and enforcement. In Cybercrime in Progress: Theory and prevention of technology-enabled offenses (pp. 175-176). Abingdon, England: Routledge.

Van Buskirk, J., Naicker, S., Bruno, R., Burns, L., Breen, C. & Roxburgh, A. (2016). Drugs and the Internet, Issue 7, October 2016.

Van Buskirk, J & Naicker, Sundresan & Roxburgh, A & Bruno, Raimondo & Burns, Lucinda. (2016). Who sells what? Country specific differences in substance availability on the Agora dark net marketplace. International Journal of Drug Policy. 35. pp.16-23.

Van Buskirk, J., Roxburgh, A., Naicker, S. and Burns, L. (2015). A response to Dolliver's "Evaluating drug trafficking on the Tor network". International Journal of Drug Policy, 26(11), pp.1126-1127.

Sivakorn, S., Polakis, J., & Keromytis, A.D. (2016). I'm not a human : Breaking the Google reCAPTCHA.

What is the Jupyter Notebook?. (2019). Retrieved from https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html

Zetter, K. (2013). How The Feds Took Down The Silk Road Drug Wonderland. *Wired*. Retrieved from https://www.wired.com/2013/11/silk-road/
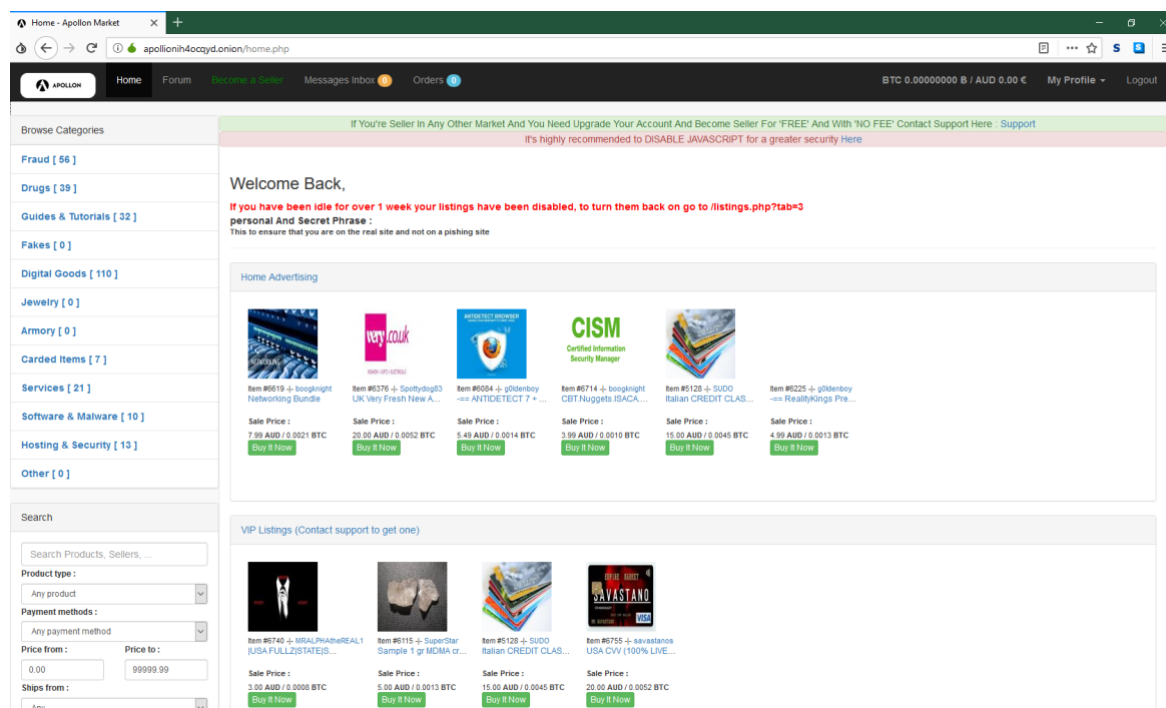
Appendix

Darknet Market Example: Apollon

Apollon market is a small and relatively new marketplace established in 2018. It is a classic escrow (payment held by third party, in this case the marketplace, until the item has been delivered) market which has both physical and digital goods listings. As of the 21st of February 2019, there are 270 items currently listed. Physical goods account for 32.59% of total listings (n = 88). While digital goods account for 67.41% of total listings (n = 182). The market's administration team claims to have over 30,000 registered users. Scrapes indicate this to be much lower. As of July 2018, the marketplace has banned Fentanyl listings.

Features:

- **Security Features:** Two-factor Authentication, Escrow, Forced PGP
- **Number of Unique Listings:** 270
- **Crypto Support:** Bitcoin
- **Language Support:** English
- **Vendor Bond:** .0333 BTC (AUD$183.17[6])
- **Insurance Bond:** $100

Image 4: Example of Apollon Main Page

---

[6] As of 21st of February 2019.

Image 5: Example of Apollon Main Page (HTML format)

This is the format collected by the crawlers described below. The crawlers currently do not collect images as it is not considered in scope in any current data analysis work.

Page, Product, and Vendor Crawler(s)

Excerpt 1: Example from Product Crawler:

```
def wget_product(directory, cookie, url, product):
    """
    Variables:
        Directory - The location to save the downloaded html file
        Cookie - Website cookie to avoid DDoS and account login protections
        URL - Website URL portion to be downloaded
        Product - Listing ID for the product page to be downloaded.
(Denoted
        by server on marketplace)
    """
    os.system('torify wget --load-cookies ' + cookie +
            ' "http://' + url + '.onion/' + product +
            '" -O "' + directory + product.replace('listing.php?ls_id=',
'') + '.html"')
```

The above excerpt shows the main function that crawls the darknet marketplace pages. The function itself only downloads a single page, and must be implemented such that the entire site is covered. Once the topology of a site is known (which can be manually verified), the function should be looped throughout all pages to give a complete snapshot of a site. The excerpt also employs command-line functionality to implement the *torify* and *wget* functions which are vital to downloading a web page through the *TOR* browser service.
The process can also be implemented in a two-stage verification method to ensure pages are successfully downloaded, with a re-download occuring for initially unsuccessful calls.
The usage of variables are defined within the code box.

Page, Product, and Vendor Scraper(s)

Excerpt 2: Example from Product Scraper:

```
try:
    with open(location + file, 'r') as html_file:
        soup = BeautifulSoup(html_file, 'lxml')
        try:
            product = (soup.select('div.col-sm-12')[i]
                        .get_text()
                        .replace('\n', ''))
            price_in_aud = ''
            price_in_btc = ''
            vendor = (soup.select('div.col-sm-3 > small')[i]
                        .get_text()
                        .split()[0])
```

The above excerpt from the product scraper shows an implementation of importing the downloaded html files into a searchable and editable object using *BeautifulSoup*. Once the page is converted, a search can be performed on the object to isolate and collect the desired data. The initial process development tedious as the search functions must be manually tested and verified before the implementation of an automated script. Scapers must be regularly monitored as subtle changes to the website and their data displaying methods can change the location of the data in question. The collected data is then exported to a CSV file that can be used in analysis. If the scraper fails to extract any of the information for each record, the entire record is skipped and the process repeated for the next record, as shown by the use of the second "*try*" clause.

Most scrapers can be implemented to use the same rough heuristic method, although this will change based on the placement of data within the html file.

Glossary - External Functions Described in Crawler & Scraper Code

os.system[7] - Executes a shell command (shell specific to the Operating System used)

torify[8] - Functionality that results in any proceeding functions to be executed using the TOR network. This allows functions that require internet access to access the TOR network. More commonly known as a TOR 'wrapper' as it encloses any specified calls in the TOR service.

wget[9] - Functionality included in GNU that retrieves files using common and widely-used Internet protocols (including HTTP, HTTPS).

BeautifulSoup[10] - An additional python package that translates the initial HTML file into a traversable data object that can be searched.

---

[7] https://docs.python.org/2/library/os.html

[8] https://linux.die.net/man/1/torify

[9] https://www.gnu.org/software/wget/

[10] https://www.crummy.com/software/BeautifulSoup/bs4/doc/