Tor marketplaces exploratory data analysis: the drugs case

Alessandro Celestini
Institute for Applied Computing, IAC-CNR,
Via Dei Taurini 19, Rome, Italy
a.celestini@iac.cnr.it

Gianluigi Me
CERSI- Luiss Guido Carli University
Viale Romania 37, Rome, Italy
gme@luiss.it

Mara Mignone
RISSC, Research Centre on Security and Crime
via Casoni 2, 36040 Torri di Quartesolo (VI), Italy
mara.mignone@rissc.it

Abstract

The anonymous marketplaces ecosystem represents a new channel for black market/goods and services, offering a huge variety of illegal items. For many darknet marketplaces, the overall sales incidence is not (yet) comparable with the correspondent physical market;however, since it represents a further trade channel, providing opportunities to new and old forms of illegal trade with worldwide customers, anonymous trading should be carefully studied, via regular crawling and data analysis, in order to detect new trends in illegal goods and services (physical and digital), new drug substances and sources and alternative paths to import socially dangerous goods (e.g. drugs, weapons). Such markets, based on e-commerce retail leaders model, e.g. Amazon and E-bay, are designed with ease of use in mind, using off-the-shelf web technologies where users have their own profiles and credentials, acting as sellers, posting offers, or buyers, posting reviews or both. This lead to very poor data quality related to market offers and related, possible feedback, increasing the complexity of extraction of reliable data.

In this paper we present an approaching methodology to crawl and manipulate data for analysis of illicit drugs trade taking place in such marketplaces. We focus our analysis on AlphaBay, Nucleus and East India Company and we will show how to prepare data for the analysis and how to carry on the preliminary data investigation, based on the Exploratory Data Analysis.

***K*eywords** tor, marketplaces, dark web, exploratory data analysis

1 Introduction

More than 20 anonymous marketplaces [1] are currently active in the *dark web*, a share of the so-called deep web. Specifically, anonymous marketplaces are mainly implemented as hidden services in The Onion Router (Tor) network, an overlay network providing anonymity to its users. The Tor network reroutes user's connection through multiple anonymous servers (onion routers), masking the original IP address of the user. Anonymity of both administrators and participants is ensured by the use of different technologies: electronic payments are carried out through the use of the virtual *cryptocurrencies* (Bitcoin is the most popular of near 100 cryptocurrencies), *Tor* is used for network

[1] On 9-9-2016, www.deepdotweb.com

communication and the *Pretty Good Privacy (PGP)* cryptosystem is often used to exchange e-mails. Darknet marketplaces operate on the same model as eBay, with three main actors: *vendors*, *buyers* (both *participants* of the Marketplace) and *market administrators*, taking a 5-10% cut of each sale and providing the web-platform, the escrow service and the basic rules under which all market participants must operate. Active running marketplaces are continuously vatying in number, due to scams or taking downs, offering a plethora of illegal goods services, e.g. weapons, drugs, pharmaceuticals, C&C systems, digital identities, counterfeited goods, contraband. In particular, according to the Global Drug Survey [2], online markets still account for a small share of illicit drug sales, although they are growing fast and turnover has risen from an estimated \$15m-17m in 2012 to \$150m-180m in 2015.

In spite of both the escrow and feedback systems (like those in Amazon and eBay), where buyers rate their purchases and leave comments to help to choose a trustworthy supplier, scams are not rare between vendors and buyers, and among administrators and users (both vendors and buyers): in this case, administrators quit&close the marketplace (so called *exit scam*), frauding the users by keeping their escrowed money, as in the case of Evolution marketplace, disappeared in March 2015 with \$12m-worth of customers Bitcoin. The frauds enabling factor, as in the most cases of digital and real life, is based on the basis of social engineering techniques, relying on the trustful, sometimes naive, behavior of the ICT user, which can easily access TOR and marketplaces due to popular *GUI's look and feel*. In fact, the success of darknet marketplaces has been facilitated by the development of usable interfaces to anonymous networks, (e.g. the "Tor browser bundle") that made it easy for anybody to browse the Internet anonymously. Moreover, anonymous marketplaces integrate many structural features of popular web marketplaces, providing a searchable listings of products for sale where buyers are leave feedback on their purchases, regardless of the trustworthiness of the comments posted by other users. Consequently, many vendors are taking the risks of shipping internationally and we can expect this trend to continue moving forward with the expansion of cryptomarkets, as shown in [1]. Based on the above-mentioned facts TOR marketplaces can be considered outsiders in the market drug, offering a new sales channel very hard/expensive to cope with in the single transaction: therefore, the importance of the preventive approach by monitoring the activities on cryptomarkets, enables the early warning on changes in the evolution of drug use in different countries and prepares the development of prevention programs that target the drugs being sold.

2 Related Works

Several recent works focused on the study of Internet organised crime[2, 3], in particular the study of on-line drugs marketplaces[4, 5] has become quite popular, e.g. focusing on the analysis of vendors' behavior [6]or identifying new challenges for LEAs [7].

Soska and Christin in [1] present an analysis of the anonymous marketplace ecosystem evolution. Their study is a long-term measurement analysis: in more than two years, they collected data from 16 different marketplaces, without focusing on a specific products' category. With respect to this work, our study focuses on illicit drugs trade, setting a short-term analysis on a reduced set of marketplaces. The results of their study suggest that marketplaces are quite resilient to law enforcement take-downs and large-scale frauds. They also evidence that the majority of the products being sold belongs to the drugs category. Several research works about anonymous marketplaces focuses on Silk Road, in particular on the drug selling: in [8] and [9] the authors present interesting studies about Silk Road user's experiences, in both cases the analysis concerned drug purchasing. In [9] they monitored and observed the market's forum for four months and collected anonymous on-line interviews of adult 'Silk Road' users. In [8] they present a single case study, reporting the motivations, experiences and usage of the website of a single 'Silk Road' user. Christin in [10] presents a comprehensive analysis of the Silk Road marketplace. For the study the author gathered and analyzed data over eight months, obtaining a detailed picture of the type of goods being sold. The analysis, among the others results, shows that Silk Road was mainly used as a market for controlled substances and narcotics. In [11] and [12] a similar study has been done for Silk Road 2, a new marketplace launched to replace Silk Road after its shut down in October 2013 by the FBI. The main objective of these works is to compare the two markets, the old one e the new one, in particular

[2] https://www.globaldrugsurvey.com/wp-content/uploads/2016/06/TASTER-KEY-FINDINGS-FROM-GDS2016.pdf

the authors analysed the structure of the two markets and the typology of their users. The authors of [13] analyze consumer motivations for accessing marketplaces and the factors associated with their use. They recruited an Australian national sample that was asked about purchasing substances from dark net marketplaces and the reasons for doing so. In [14] the authors present an overview of the Canadian illicit drug market, providing information about how vendors are diversifying and replicating across marketplaces. Their study gives an insight into the structure and organization of distribution networks existing online. They collected data about listings and vendor profiles on eight anonymous marketplaces between August and September 2014.

3 System Description

The typical structure of anonymous marketplaces offer a list of products and their individual pages where authorized vendors can set up a virtual shop and place listings. Items for sale are organized in categories and subcategories, the organization vary from market to market, but there is large agreement on mains product categories (e.g. drugs, weapons, frauds). Usually it is possible to search products for sale both by product categories and by keywords, but the last option is not always available. Buyers and sellers are able to leave feedback about their transactions, these are usually composed of a rating (e.g., good/bad or a value between 0 and 5), a comment and the obfuscated user' nickname who leaves the feedback. Such information are used to construct users' reputations inside the market both as sellers and as buyers. The main difference with surface market is the regulation of market access: users accounts of anonymous marketplaces are needed not only to carry out transactions, but they are required to access the market itself, this isn't true for surface markets.

For our analysis we focused our attention on three marketplaces: AlphaBay, Nucleus and East India Company. The Nucleus market[3] was established on November 24, 2014, the AlphaBay[4] market was established on December 22, 2014 and the East India Company[5] market was established on April 28, 2015. Two of these three markets are not anymore active, only AlphaBay is reachable and active, the shut down of Nucleus and East India Company have been reported as cases of exit scam by market administrators. In such cases, administrators lock users' funds on the market, taking as much money as possible just to shut down the market soon after. The shut down of Nucleus was quite unexpected because it was one of the major market with AlphaBay and largely trusted by users.

For the creation of our dataset we crawled regularly the three marketplaces, visiting each market every month for four months. We focused our attention on illegal drug trade, collecting all items' pages concerning drugs and their related images. Gathered data form a dataset of 36 GB in size, each crawl of each market ranged from 1,682 to 18,831 pages/images, and crawling time took from several hours up to few days. For each market we designed and implemented a custom spider using the $scrapy$[6] framework, spiders are configured to use Tor [15] in order to have the capability of reaching anonymous marketplaces. Fig. 1 shows the logical workflow of our system for the collection, elaboration and analysis of anonymous marketplaces data.

3.1 Spiders

The crawling of anonymous marketplaces raises several challenges, such as the login and session cookies management, the avoidance of undesired actions execution and the monitoring of requests frequency. As described below, we designed and implemented our spiders trying to face all these challenges.

For each marketplace we created an account, which is mandatory to enter any marketplace. Since our spiders are able to authenticate them-self and to manage session cookies, the crawling of the marketplace after authentication is automatic. In particular, only the login phase is supervised, indeed in most marketplaces a CAPTCHA is used to protect login form, and often a double check is required by the market web site. A first CAPTCHA is used for bot exclusion and protects the login page, a second CAPTCHA protects the login procedure. We manually resolve CAPTCHA

[3] nucleusoitxmebfx.onion

[4] pwoah7oh4jlgdwri.onion

[5] g4c35ipwiutqccly.onion

[6] Scrapy: A Fast and Powerful Scraping and Web Crawling Framework `http://scrapy.org`

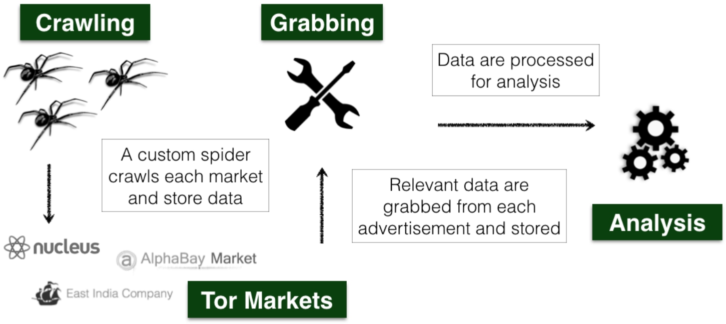
Crawling
Grabbing
Data are processed for analysis
A custom spider crawls each market and store data
Relevant data are grabbed from each advertisement and stored
Analysis
nucleus
AlphaBay Market
East India Company
Tor Markets

Figure 1: System workflow

and provides to spiders the session cookies obtained. After authentication each spider is able to manage session cookies and automatically crawl each market. Concerning requests frequency, in order to not overwhelm marketplaces with requests, we fix a random interval for each request to each marketplace. Thus, spiders request pages with different waiting times between each request, we also deny the possibility of sending concurrent request to marketplaces, clearly these precautions increases the crawling time for each market. Such settings are natively supported by $scrapy$, and we resort on them for spiders' configuration. Finally, in order to prevent undesired actions execution, such as adding items to cart or sending messages, each spider is designed to request only URL of interest. Specifically, we crawl only listings' pages and items' pages. Spiders scrape listing pages to collect items' pages URL, but they store only item's pages, together with items' images.

In order to check the outcome and accuracy of each crawling we proceed as follows: first, we manually check the number of items for sale available in each section of each market. Then, for each crawl we record all the url of items for sale in each section of each market found by spiders. Finally we count the number of item's pages gathered by each spider. If these three values coincide we are reasonable sure of the completeness of our crawling.

3.2 Data Grabbing and Reduction to Table

From each item's page we extract the information necessary to fulfill the table fields reported in Table 1. To grab such information we take advantage of the HTML structure of each page. A generic HTML page is enclosed in a main <`html`> tag and divided in two sections identified by the tags <`head`> and <`body`>. Tags inside the <`body`> contain information we look for, the data we want to extract are enclosed in different tags at different level of the HTML structure. Every single marketplace has a different tree structure of HTML pages, thus we implemented a customized parser for each marketplace:hence, the related parser identifies the nodes within the Document Object Model (DOM), through a Xpath selector.

The objective of data grabbing is to *reduce* items advertisement to *tuple*, containing the relevant field for the analysis, as shown in the following Table 1. Actually, the most valuable fields, with regard to the outcomes presented in this paper, are related to shipments, as the <`shipping_option, ships_from, ships_to`>

3.3 Why the user feedback field has not been included

While several characteristics of electronic markets serve to facilitate trade, online transactions also involve greater uncertainty and increased opportunities for fraud. Unlike buyers in traditional settings, online customers are physically unable to inspect products for sale and must rely on pictures and descriptions provided by the seller (information asymmetry). Buyers can determine the quality of a product only when they receive it, after the purchase has been made, thus sellers have less incentive to provide high quality products (e.g. lemons market). Electronic marketplaces like eBay have attempted to reduce fraud and assess sellers' reliability by allowing participants to post feedback about their transactions (signal). Tor marketplaces are characterized by higher information asymmetry: illegal trade of goods/services reinforces information asymmetry due to the poor reliability of the criminal activity of the vendor, increasing the fraud risk. Online legal markets (e.g. eBay)

Table 1: Extracted fields

Field Name	Field Type	Field Name	Field Type
name	text	product_type	text
vendor_name	text	description	text
quantity_g	decimal	product_review	text
price_bc	decimal	refund_policy	text
marketplace	text	shipping_options	text
index	int	quantity_t	text
escrow	tinyint	price_t	varchar
payment	text	ships_from	text
date_visti	date	ships_to	text
raw_material	tinyint	url	text

try to protect customers with signals like vendor reputation (feedback), but such mechanisms do not completely protect from fraud. Tor Marketplaces replicate these mechanisms and reinforce them by offering escrow services and sometimes, by denying the *finalized early* (pay the goods before shipping) option. However, several examples show that even vendors with high reputation scores decide for exit scam (e.g. 9THWonder). In conclusion, non-negligible fake feedback reviews rate, exit scams and out-of-marketplace finalization of transaction pose several doubts to the reliability of feedbacks as a metric to estimate vendors sales volume and her/his reliability, considering that every single purchase can be a black swan event.

4 Data Preparation

The marketplace mechanism to place offers makes the overall raw dataset quality very poor. In fact, every vendor can place his offer regardless of almost any input data check, leading to the creation of empty or duplicated offers (in the raw dataset), mistakes in key-text fields, and non-uniform metrics/currencies. Moreover, the offers can be misplaced in the different categories, leading to mismatching the overall amounts of offers. Hence, for the above mentioned reasons, the data preparation phase assumes a strategic role in order to achieve correct analysis results. Former phase relies on *data scrubbing*, also called *data cleansing*, which is the process of correcting or removing data in a dataset that is incorrect, inaccurate, incomplete, improperly formatted, or duplicated. The result of the data analysis process not only depends on the algorithms, it also depends on the quality of the data, that is why the accuracy of the data scrubbing process is crucial after data collection. Hence. in order to avoid dirty data, the dataset should be characterized by *correctness, completeness, accuracy, consistency and uniformity*. Dirty data (e.g outliers) can be detected by applying statistical data validation methods and by parsing texts, deleting duplicate values and missing data contribute to bad analysis results. After the collected dataset has been loaded into the data preparation database, containing all the raw data to be manipulated, some statistical methods have been applied to find values that are unexpected and thus erroneous, even if the data type match but the values are out of the range, it can be resolved by setting the values to an average, Median, and Range Constraints.

Resuming, the data preparation phases applied to the raw dataset are:

- *Data cleansing and wrangling* represents a core process in this work to prepare data, which are directly inserted by vendors without strong data validation constraints. In this phase incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data are replaced, modifyied or deleted. In fact, offer attributes as <`weight`>, <`price`> and <`description`> can be found in different metrics (e.g. ounce, gram or BTC, LTC, USD).
- *Data Transformation*, where the mapping of the data from its given format into the format expected by the analysis tool. This phase includes value conversions or translation functions, as well as normalizing numeric values to conform to minimum and maximum values;
- *Data Exploration*, applying Exploratory Data Analysis (EDA) which represents a critical part of the data analysis process because it helps us to detect mistakes, determine relationships and tendencies, or check assumptions;

- *Modeling*, which represents the most important stage in the analysis with selection, application, evaluation of results and possibly model tuning;
- *Visualize and interpret results*, which represents the key phase to let the results understandable, improving their overall impact.

Data cleansing, wrangling and exploration phases have been implemented by the use of Rapidminer [7].

The above mentioned data collection and preparation processes have been carried out in order to answer the following questions:

1. How many offers (overall/Novel Psychoactive Substances)?
2. How many vendors?
3. Where do they ship from?(Possible drug stocking place)
4. Which is the world share distribution?
5. Which is the European countries' distribution map?
6. Which market share of the overall categories is holding the half of offering vendors?
7. How many possible scams?
8. Which vendors are cross-sellers on categories?

The following preliminary indicators fulfill the above requirements, putting EDA in practice and supporting for data cleansing and wrangling:

- *Offers*: the overall number of offers considered (1)
- *Number of Vendors*: the overall number of vendors (2)
- *Median*: the value of the number of offers of the vendor splitting halfway into the set of vendors
- *Offers to median*: overall number of offers to median value
- *Market polarization* (%):
$$MP = \frac{\text{offers to median}}{\sum \text{offers}}$$
- *Possible scam* (%): the ratio between the number of vendors offering one or two items and the total number of vendors (7)
- *European share(%)*: the ratio between the European offers and the total amount of offers

Although in the following section the authors will present some sample charts, the overall results analysis is out of the scope of this paper.

5 Data Analysis

The obtained dataset is ready for EDA, providing results (presented in this paper only in a sample) in the form of both uni and multi-variate graphical statistics. We used EDA to extract important variables, discover underlying structure in order to further develop appropriate models. One limit of this methodology is related to the reliability of the offers, since the only way to verify if an offer is a fake or a scam is to proceed with the purchase of the item and wait the shipment. During data exploration we recognized vendors with high-volume of offers in different markets, we can assume that these are not fake vendors (although we can never avoid the scam), while nothing can be said about vendors with low-volume of offers. Hence, we assumed that fake offers can be placed by vendors with less than 3 different offers, as shown in Section 4.

As shown in Figure 2, the most recurrent offers on the Eastindia, Nucleus and Alphabay markets are in the cannabis category (natural, not synthetic), followed by the ecstasy and stimulants categories. This ranking is repeated individually for the three markets, showing which are the most requested substances in each market.

[7] https://rapidminer.com

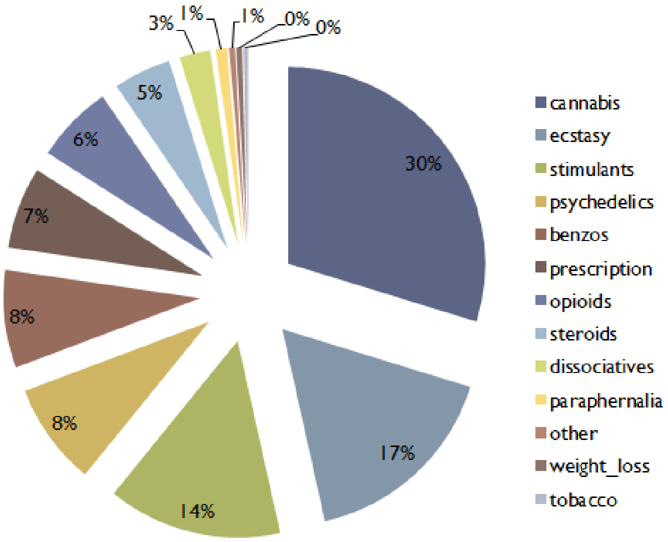

Figure 2: Overall Shares of Drugs market

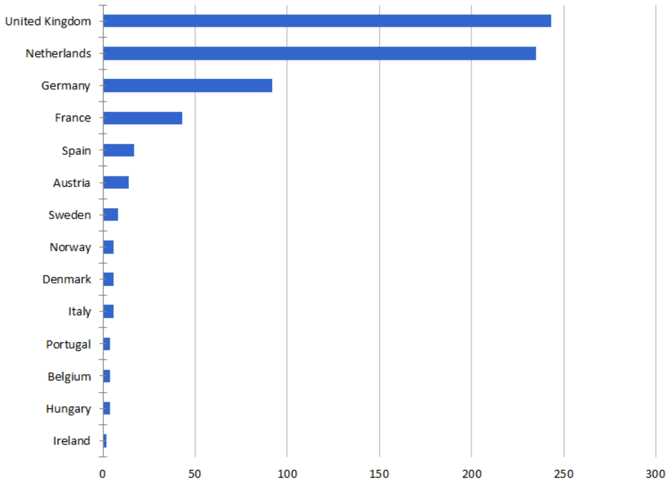

Figure 3: Eastindia offers by origin country

The Figure 3 shows, for Eastindia market, which are the top European countries labeled as shipping departure places for drugs. These data are mostly reliable due to the fact that shipping expenses change with respect to the destination continent/country. Moreover, from these data we can infer what is the country where the item is located, while nothing can be said regarding vendor's location (which in many cases resides in the shipping source).

Finally, in Figure 4 small bubbles indicate low number of countries selling items in the related category. Showing that $cannabis$, $stimulants$, *benzodiazepines* and *opioids* offers are widespread all around the world, probably related to facilitated access to the substances, while $prescription$, $steroids$, *tobacco* and $weightloss$ are concentrated in few countries, probably related to facilitating legal framework. Ecstasy represents an anomaly with a very high number of offers concentrated on average number of countries, but most of the offers are from the Netherlands.

6 Conclusions

Illegal trading sites on the Darknet represent cross-cutting crime enablers. In particular, recent studies [6] showed that drug market represents the biggest offers share in every TOR marketplace, due to

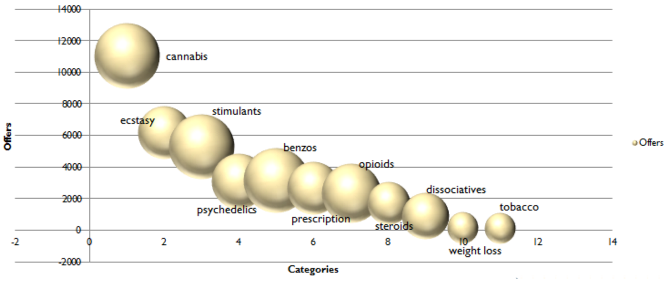

Figure 4: Categories size vs cross market vendors

reinforced motivation provided by the greater anonymity afforded by the Tor Network. The huge size of the listings enables multiple trading opportunities, e.g. retailing (dealer-to-customer transactions). Since the impact of these new dynamics in the overall drug market (or in some categories/segments) is not yet completely clear, monitoring trends and behaviors represent a crucial cornerstone to cope with this phenomenon, as confirmed, e.g., by the marketplace take-downs (e.g. operation Onymous), not representing an effective response to this phenomenon, since the underground community regularly recovers to some degree, showing resiliency to this countermeasure. In order to determine characteristics of vendors, substances, countries and prices and to determine significant differences in trends among all these drug offers features, we presented a methodology to automatically surf and prepare data from offers on Tor marketplaces in order to target a preliminary EDA on drug offers data and, consequently, providing hypothesis to drive further refined analysis of the phenomenon. In fact, our ultimate, ambitious objective is to understand the underlying *modus operandi* and to detect criminal mechanisms representing the base of the illicit drugs trade phenomenon taking place in the Tor marketplaces.

7 Acknowledgement

This work has been supported by the Drug Prevention and Information Programme of the European Union under the EPS/NPS Project -Enhancing Police Skills concerning Novel Psychoactive Substances, with the financial support of the European Commission Targeted call on cross border law enforcement cooperation in the field of drug trafficking DG Justice/DG Migrations and Home Affairs (grant JUST/2013/ISEC/DRUGS/AG/6429). This research work has been partially funded by IANCIS[8], a EU ISEC project.

References

[1] Kyle Soska and Nicolas Christin. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 33–48, 2015.

[2] The internet organised crime threat assessment (iocta). Technical report, EUROPOL, 2015.

[3] Kurt Thomas, Danny Huang, David Wang, Elie Bursztein, Chris Grier, Thomas J. Holt, Christopher Kruegel, Damon McCoy, Stefan Savage, and Giovanni Vigna. Framing dependencies introduced by underground commoditization. In *Workshop on the Economics of Information Security*, 2015.

[4] Mara Mignone and Elisabetta Bosio. Criminological analysis of the nps market. Technical report, RISSC, 2016.

[5] Luigi Laura and Gianluigi Me. Searching the web for illegal content: the anatomy of a semantic search engine. In *International Conference on Global Security, Safety, and Sustainability*, pages 113–122. Springer, 2015.

[8] `http://www.iancis.eu/`

[6] Diana S Dolliver and Jennifer L Kenney. Characteristics of drug vendors on the tor network: A cryptomarket comparison. *Victims & Offenders*, pages 1–21, 2016.

[7] Julia Buxton and Tim Bingham. The rise and challenge of dark net drug markets. *Policy Brief*, 7, 2015.

[8] Marie Claire Van Hout and Tim Bingham. 'Silk Road', the virtual drug marketplace: A single case study of user experiences. *International Journal of Drug Policy*, 24(5):385 – 391, 2013.

[9] Marie Claire Van Hout and Tim Bingham. 'Surfing the Silk Road': A study of users experiences. *International Journal of Drug Policy*, 24(6):524 – 529, 2013.

[10] Nicolas Christin. Traveling the silk road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 213–224, New York, NY, USA, 2013. ACM.

[11] Diana S. Dolliver. Evaluating drug trafficking on the tor network: Silk road 2, the sequel. *International Journal of Drug Policy*, 26(11):1113 – 1123, 2015.

[12] Rasmus Munksgaard, Jakob Demant, and Gwern Branwen. A replication and methodological critique of the study "Evaluating drug trafficking on the Tor Network". *International Journal of Drug Policy*, pages –, 2016.

[13] Joe Van Buskirk, Amanda Roxburgh, Raimondo Bruno, Sundresan Naicker, Simon Lenton, Rachel Sutherland, Elizabeth Whittaker, Natasha Sindicich, Allison Matthews, Kerryn Butler, and Lucinda Burns. Characterising dark net marketplace purchasers in a sample of regular psychostimulant users. *International Journal of Drug Policy*, pages –, 2016.

[14] J. Brosus, D. Rhumorbarbe, C. Mireault, V. Ouellette, F. Crispino, and D. Dcary-Htu. Studying illicit drug trafficking on darknet markets: Structure and organisation from a canadian perspective. *Forensic Science International*, 264:7 – 14, 2016. Special Issue on the 7th European Academy of Forensic Science Conference.

[15] Roger Dingledine, Nick Mathewson, and Paul Syverson. Tor: The second-generation onion router. Technical report, DTIC Document, 2004.