Analysis of Hacking Related Trade in the Darkweb

Othmane Cherqi [1] [2] , Ghita Mezzour[1] , Mounir Ghogho [1] , Mohammed El Koutbi[2]
[1] Universit Internationale de Rabat, Faculty of Informatics and Logistics, TICLab
[2] Mohammed V University, ENSIAS
Emails: {othmane.cherqi, ghita.mezzour, mounir.ghogho}@uir.ac.ma, koutbm@gmail.com

***Abstract*—The non-referenced web is estimated at five hundred times the size of the surface web. Darkweb represents about 6% of the non-referenced web and includes all kinds of delinquency: drug trafficking, counterfeiting, hacking market etc. Several studies about Darkweb marketplaces have been carried out, focusing mainly on the analysis of drug trafficking or the extraction of products sold in different forums. To the best of our knowledge, by the time of this study, no work has yet been done to analyze hacking trade business.**

Most hackers rely on malwares and softwares offered in different cyber markets to commit their attacks. The main objective of our work is to present an exploratory analysis of the illegal trade that is developing in this marketplaces in order to have a clear idea of threats that may harm individuals, industries and organizations. Through this work, we have been able to give a clear insight on the hacking market. The main motivation of sellers in this market is profit making, in fact this market generated over 26 million USD during the period studied. Product accessibility is also an alarming factor: 85% of the products offered do not exceed 150 USD, making cyber crime accessible to all. Finally, one cell controls almost the entire market, indicating the presence of a well-organized infrastructure. The results of this analysis are discussed below.

I. Introduction

Darkweb Marketplaces are online markets generally hosted as Tor Hidden Services that serve typically for irregular products trade. Identifying and analyzing products being bought and sold in such platforms can have important implications specially for threat intelligence and security related application. The collection of such content is also important for studying and understanding the behavior of sellers and consumers presence in these online communities. What makes the Dark web difficult to study is its volatility. Indeed, these websites are not well referenced and their access requires a minimum of expertise. In addition to this, these sites are regularly taken out of service by governments or ”White Hackers ”or simply change URLs to avoid being spotted. Several works have been done in this direction to develop the most efficient tools to collect data from Darkweb's forums and markets and analyze it. The study presented in [1], focused on scrapping Darkweb forums in order to identify eventual threats within hacker's exchanges. The research described in [2] was conducted to obtain information on the types of products sold on Agora (one of the biggest Darkweb markets) i.e. narcotics, counterfeiting and weapons. Of course, the information presented on these platforms is not always structured and often requires prepossessing after harvesting. Nunes, Eric, et al. proposes a system that collects information from different forums and markets on surface web and dark web and determines, using NLP, whether this collected information is hacking-related or not. However, no exploitative analysis was done targeting the trade of malicious hacking. A neglected area in the field of Darkweb marketplaces is the analysis of sold items and most influential vendors which, we believe, may give valuable insight about general cyber crime trend. This exploratory data analysis consists on mining the hacking transactions on major Darkweb marketplaces during there period of activity. We use scraped data from Agora, Aplhapbay, Silkroad2 and Nucleus from 2013 to 2015, which represents more than 900k announcement. This paper makes several contributions to understanding how hacking related items' trade is managed on Darkweb marketplaces. First, each transaction is categorized based on whether it's hacking related or not. Using supervised learning techniques and manual labeling ground truth, we aggregated all the hacking related ads into one category that we called ”Hacking related” and we labeled other announcements as ”Not Hacking related” as shown in Table I. Next, we give a general economic impression of the market. In other words we study the prices charged on the market, the profit made by the sellers and the total value of the products offered. Finally, we investigate whether there is some group of individuals who control the market, and on what basis do users choose with whom to make their transactions. The rest of the article is organized according to the following plan: Section II gives an overview of the data used and section III presents the preprocessing steps that have been done to the data. In section IV we explain the process followed to label and extract hacking

TABLE I: Example of Items' Labeling

Given Label	Item Description
Hacking related	- Mac & windows address changer. - ANONYMITY TESTING SERVICE. - How To Setup a Botnet Guide . - DDOS ATTACK SERVICE.
Not hacking related	- Counterfeit $20 USA currency. $200 worth. 10 pre cut bills Free shipping. - Hublot big bang Replica. - 5 EasyTouch 1cc Syringes 28G x 1/2. - Make $400+ a Week with Fiverr.

978-1-5386-7848-0/18/$31.00 ©2018 IEEE

related items from the collected data. Section V concerns the results of the exploratory data analysis conducted on hacking ads. On section VI we give an overview of related work and finally we give limitations and conclusion on section VII and VIII respectively.

II. Related Work

We focus our literature review on previous studies concerning Darkweb exploration. Crawlers are the ideal tool to get data from web pages. However, this collection is often aimed at a specific subject. Hence the importance of setting up crawlers that can be adapted to different themes, more commonly called focused crawlers [9], [8]. However, these crawlers are designed to access referenced web sites. Yet in [10] they designed a deep web oriented crawler. This research focused primarily on forums, data collection and statistical analyses to study online communities. At the end of this research they provided analyses on the different communities present in the forums collected during a certain period. The authors also discussed the different exploratory techniques used on [11]. Complementary, [3] focus not only on forums but also in marketplaces in order to scrap all information related to hacking. Also, some work similar to ours has been done, namely studying the trades present in the Darkweb. In [12], they analyzed the hacker community and card data smuggling , and in citeb4, they focused their work on mining drug and fake ID cards businesses. On our side, we not only extracted the products related to hacking from the Darkweb, but we also conducted an exploratory analysis of these products.

III. Dark Web Market Places Data set

The first section concerns the presentation of the data we used. We referred to an archive from over 80 different web markets made available for research. Since 2013, researcher Gwern Branwen has been gathering all kinds of information related to the Darkweb market: articles, customer feedback, product photos, forum discussions etc., on weekly or daily basis: depending on the availability of websites. The archive contains not only the data pasted by the researcher, but also those shared by several other researchers and dark web personalities. Most of the files in the collection are scraps of websites in raw HTML format.

We consider analyzing four of the most popular and well known marketplaces : Agora, Silkroad 2.0, Alphabay and Nucleus. We analyze the advertisements that have been published and extract all information about the offers. We were able to collect a total of 972.655 items for our analysis, starting from 7th July 2013 until 20th December 2015. By looking at the supply side of Dark web, most of the ads are not related to hacking but rather to drugs as shown in Table II (25 and 75% respectively) . However, cybercrime seems to have a significant share in terms of value of offered products (26 million USD).

TABLE II: Marketplaces Database Details (Dec. 2014 - Jul.2015)

Marketplace	**Listed items**		**Vendors**		**Value($)**	
	Tot.	**Hacking**	**Tot.**	**Hacking**	**Tot.**	**Hacking**
Agora	536K	84K (15%)	2495	792 (31%)	63M	6M (9.5%)
Silkroad	280k	61,5K (30%)	954	300 (31%)	166 M	13.4M (8 %)
Alphabay	86K	45K (52%)	800	522 (65%)	19.2 M	6.9M (35%)
Nucleus	7698	1293 (5.9%)	1243	322 (26%)	3.5 M	183K (5%)
Total	973K	192k (%19)	5492	1936 (25%)	252M	26.5M (10)%

IV. Data Preprocessing

In this section we will describe how we reduced scraped HTML files into a structured form containing the relevant insight for exploration and analysis work. Each marketplace selected has a specific HTML page structure, thus we developed a customized parser for each one. The semi structured form of HTML and the tags specific to each field allowed us to find more easily the information we needed on each page. The main information we have extracted about the products are :

- **Vendor:** Vendor's alias
- **Category:** Category item belongs to
- **Item:** Item's description
- **Price:** Cost of the article (average in case of duplicates)
- **Origin:** Seller's location (remains unreliable, as the majority of vendors wish to preserve their anonymity)
- **Destination:** Destinations vendor ships to
- **Rating:** Score given by consumers to vendors based on a stars rating scale.

The parsed pages containing ads as entered by the sellers are very noisy, thus they need to be cleaned. Moreover, in order to focus only on hacking transactions, we classify each item whether it's hacking related or not. To that end, we need to label the extracted items.

Data preprocessing phase assumes a strategic role in order to achieve correct analysis results. This is why we had to remedy all the data deficiencies mentioned above. In the following we list the cleansing work applied to the collected data:

- Words misspelling: Misspellings is very recurrent when the user is given the task of entering information.This step consisted in fixing this, for example correction of origin and destination input errors based on regular expressions and countries spelling.
- Mixed up currencies: Some products have their unit value incorrectly indicated as Bitcoin, when in fact they are in USD. This is detected based on the price comparison of the same products. Values deemed outliers are converted to USD.
- Information entered in the wrong field: here too we use regular expressions to determine if a data has been entered in the appropriate field i.e. detection of the

TABLE III: Median accuracy of each classifier

Model	Precision	Recall	F-score
Random Forest Classifier	0.6	0.46	0.74
Linear SVC	0.89	**0.94**	0.86
Multinomial Naive Bayes	**0.92**	0.93	**0.93**
Logistic Regression	0.81	0.94	0.84

currency for prices, list of destinations and origins etc.

Our work only targets products related to hacking, where the need to label these items by conducting a machine learning approach based on classification. The purpose of this classification is to accurately identify whether a product proposed in Darkweb marketplace is hacking related or not, based on the description given by the seller. We used a supervised learning approach combined with TF-IDF (term frequency-inverse document frequency) to resolve this issue.

Bag of Words and TF-IDF: The most popular approach when extracting characteristics from a text is to use the bag of word model: For each text, in this case the product description, we take into consideration the frequency of the terms while neglecting their order of appearance. Generally, we calculate Term Frequency, Inverse Document Frequency, abbreviated to tf-idf, for each term.
Once this work was done all that remained was to build and compare different models using tf-idf features matrix.

Classification Models Benchmark: The purpose was to classify the items by analyzing their description in a supervised way, and the objective at this level is to determine which supervised machine learning method is best suited to this. We want to assign each product to one of two categories (Hacking related/ Not hacking related). This is a binary text classification problem for which we use the following four models:

- Logistic regression
- (Multinomial) Naive Bayes
- Vector machine with linear support
- Random forest

We compared the performance of these models with a 5-fold cross-validation using a labelled subset of Agora marketplace from the years 2014 to 2015 provided as part of a Kaggle competition[1]. Multinomial Naive Bayes give the best classification results compared to the other models with almost 93.3% of median accuracy.
Naive Bayes Classification Results: Most predictions are on the diagonal line of the confusion matrix (true positives) with 97% true positive rate for the ”Hacking related” category and 93% true positive rate for ”Not hacking related”, as shown in Fig. 1. Table IV shows that model is performing well for categorizing items with a precision of 93% an a recall of 97% concerning ”Hacking related” category. For ”Not hacking related” category, model is getting 97% precision and 93% recall. Regarding this result we assume that the classification made by the model is reliable for the rest of the work.

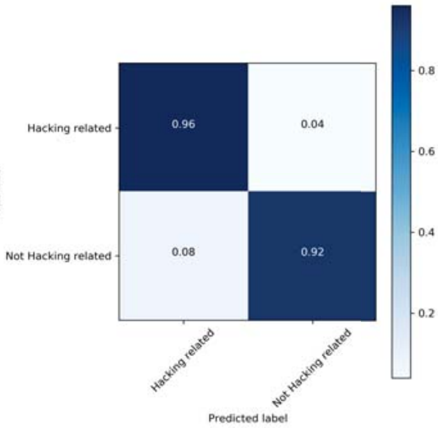

Fig. 1: Normalized confusion matrix for product classification

TABLE IV: Classification report on item labeling

	Precision	Recall	F-score
Hacking related	0.93	0.97	0.95
Not hacking related	0.97	0.93	0.95
Avg.	0.95	0.95	0.95

V. Darkweb Market Places Exploratory Analysis

The main purpose of this work is to provide some cyber threat intelligence via understanding how certain services including most common and severe malwares, zero-day threats, exploits, DDOS services etc., will occur, by answering questions such as :

- What is the share of cyber crime in the Darkweb market in general?
- Who are the main sellers?
- Is there a market monopoly on the part of some of the sellers?
- Are the ads reliable?
- Is there a real quality/price factor?
- What are the prices of the products?
- What are the most expensive products sold?

In order to study and visualize the extracted data, we have calculated some metrics defined as follows:

- *Number of items:* The total number of items offered by a particular supplier.
- *Values:* The sum of the items offered by a vendor.
- *Score:* Average of all ratings given to a supplier by consumers.
- *Average value:* Average price of items offered by vendors.

[1] https://www.kaggle.com/nicapotato/dark-net-light-in-the-dark/data

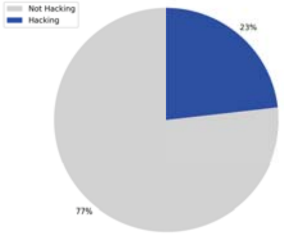

Fig. 2: Products distribution on Darkweb Marketplaces (2013-2015).

A. Market Distribution

In this section we wanted to study the contribution of hacking ads in main Darkweb markets. As one might expect we notice that cyber-delinquency represents a smaller part of dark web marketplaces compared to the ”Not Hacking related” category as shown previously in Table II, this can be explained by several factors :

– We have grouped more under categories in the category ”Not hacking related”, namely drugs, weapons, counterfeits, etc..
– Many more advertisers and ads are returned for these products because no expertise is required for this type of business unlike the product related to cybercrime. Indeed it is often necessary for hackers to follow up with customers or even to develop their products themselves.

By aggregating items related to hacking together in Figure 2, we notice the inferiority of hacking offers compared to other fields such as counterfeiting, trafficking in identity documents or narcotic drugs(grouped on category ”Not Hacking related”) 25 and 75% respectively.

B. Price Distribution

Price variation in a market is a good indicator to get an idea of supply and demand [5]. The total market size concerning hacking announcements reaches 26.5 million USD (for the period under study). Figure 3 reveals a sub-exponential distribution i.e. heavy tail for both hacking ads and other categories. Different hacking services can be found, making it a viable option for new hackers and more generally anyone interested in cybercrime. Concerning Hacking offers, around 85% of the products offered do not exceed 100 USD and most of the prices cheaper than 500 USD. This confirms our hypothesis that this kind of service is within the reach of all budgets. Compared to ID trafficking or drug sales, cyber crime seems to be the least dominant trade (10% of total market value). However some sellers make huge profits as shown in Section V-C. To have an overview of the market, we set the price limit at 500 USD for the cumulative distribution function of the prices, because almost all products offered do not exceed this limit. However, there are products for which the price significantly exceeds this margin as shown in Table V. These offers generally concern hacking as a service[2]. Indeed the most expensive items are precise services, requiring a lot of expertise and its scarcity is the main factor of high price. We find offers to rent a hacker at 15000 USD or ready for use botnets at 1345 USD.

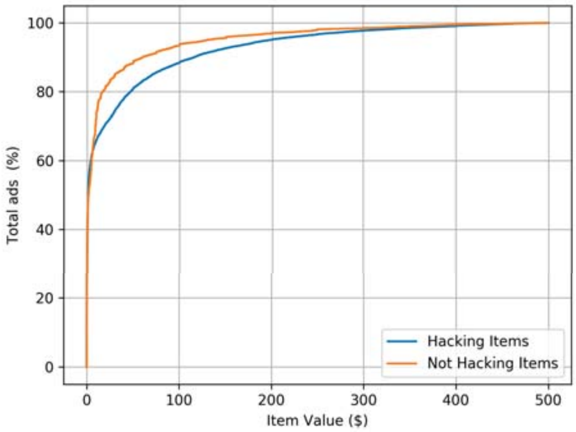

Fig. 3: Cumulative Distribution Function of Sold Items

TABLE V: Example of most expansive Items (2013-2015)

Item	**Value ($)**
Hire Hackers	15000
USA Version Ford Rolling Code Grabber Opener	1489
Get access over others Facebook Twitter and Youtube accounts!	1247
PWN PADS- Hacking pad (usually retail for $1095usd)	1016
Fully Setup Botnet - 1000 Bots	1345
Website hacking - Get passwords e-mails all database - PM ME	1623
Custom Hacking Order for PapaEmeritus	5076
iCloud Server Clone	3990

C. Seller’s Notoriety

After studying the economic trend of the cybercrime market, this section is devoted to emphasize whether there is a group-like of sellers who take advantage of all this wealth. This part also provides additional support for getting an overview of the volume of leading providers in the field of cyber criminality. Our objective is to put the finger on the distribution of the sellers, to update eventual monopolies and to see if there are indicators which would show that we deal with well organized criminals.

[2] Hacking as a service (HaaS) is the commercialization of hacking skills, in which the hacker serves as a contractor

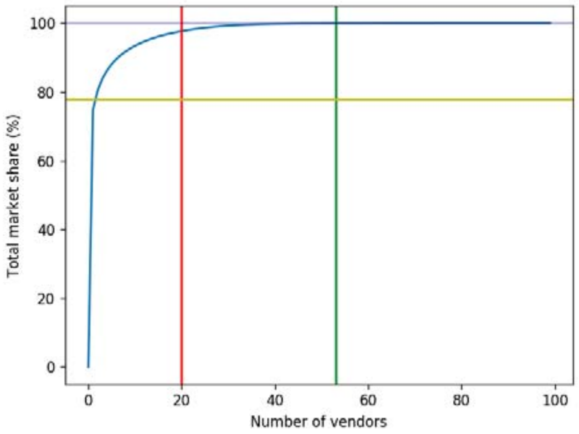

Fig. 4: Pareto Distribution of Darkweb Marketplaces

Figure 5 shows the ranking of the top 30 biggest sellers in term of product supply, and it is interesting to see how the disparity in the total value of goods offered by each hacker manifests itself. We clearly distinguish from the power low distribution of the supply volume how the market is dominated by few vendors. This suggests that there could therefore be a monopoly on the profit generated. Thus, the Pareto principle [5] emerges from the monopoly tendency of some sellers since in this case, the 80/20 rule applies pretty closely as can be seen in Figure 4. 20% of sellers have a monopoly on more than 98% of the market. Such a monopoly clearly indicates that there is a whole structure involved in managing such a market.

Such concentration clearly indicates that there is a whole structure involved in managing this cyber crime market. This is large-scale organized crime, not a small independent handymen. This phenomenon is due to the fact that we are facing an oligopolistic market, which in economics results from great demand but little supply. This means that we are dealing with a burgeoning market and there is an increasing demand for hacking products.

Table 5 gives the inventory value of Darkweb marketplaces' top 10 sellers. It's immediately evident that its not small criminality, we are talking about entities like "The Interways" with over 6 million USD of product on sale over the time of our study. This indicates that there is a significant gain generated by these traffickers. Their main motivation is financial. We deduce that the sellers operating in this kind of market are neither hacktivists nor militants, they are mainly traders whose aim is to earn money.

D. Understanding Consumer Behaviour: Rating System

One of the main characteristics of e-commerce on the Internet, and in Darkweb more particularly is the anonymity of stakeholders. Thus the only indicator on the reliability of a product is the review left by consumers. In this section, we try to understand consumer behavior by analyzing their feedbacks on products offered. Figure 6 is interesting because it shows that users rate the products that have satisfied them and do not hesitate to give bad feedback to scammers . This allows to have an idea about the seller before dealing with him.

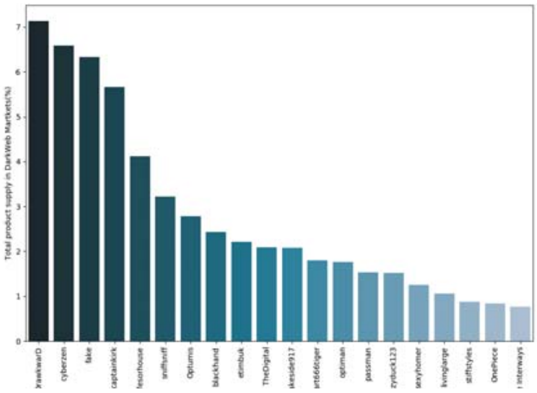

Fig. 5: Hacking ads top Vendors

TABLE VI: Darkweb Marketplace Top Vendors (2013-2015)

Vendor	Total Items	Supply Volume ($)	Item avg. Price ($)
The Interways	1608	6 062 738	3770.36
Oyasis	733	862 805	1177.09
Profesorhouse	8686	817 561	94.12
TheProfessionals	635	594 314	935.93
DrawkwarD	15042	540 427	35.93
The g0dfather	12	364 486	30373.88
PerfectScans	703	259 194	368.70
GreenDawg	155	248 000	600.00
Takri	387	246 600	637.21
technohippy	1350	149 832	110.99

It is also clear that the rate of good feedback is much higher than bad feedback. This means that most advertisers offer products in line with the ads they make. This suggests that there is more to be gained by being "honest" than by scamming people.

The most striking result to emerge from 7 is that the higher the price the higher the rating. This confirms the hypothesis made in section V, according to which, there would exist a quality/price ratio in the products offered. However Pearson correlation coefficient shows that there is no real correlation between prices and consumer ratings. Indeed we see that a large part of the products whose price is not very high have been well rated.

VI. Limitations and Future Work

In this section, we discuss the limitations of the proposed analysis of hacking trade in the Darkweb. We suggest possible

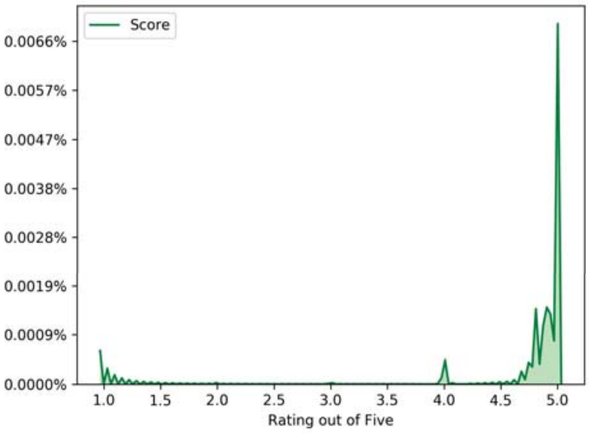

Fig. 6: Rating distribution

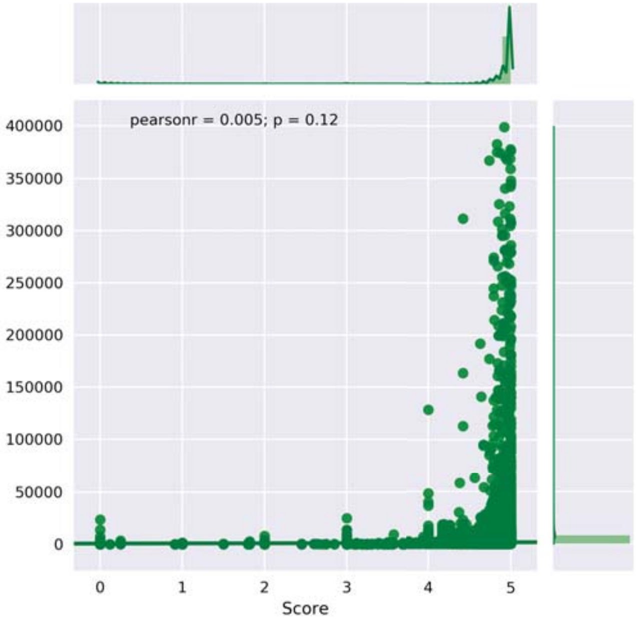

Fig. 7: Rating Variation According to Items' Prices

improvements and opportunities. We believe that a major point that is still not dealt with is the use of labeled hacking ads for cyber threats forecasting. In other words, more in-depth analysis of products sold can be made to find an potential correlation with the incidence of cyber-attacks. Also in the same sense, we plan to extend this analysis to hacking forums. This would make it possible to highlight the existence of a network that straddles the two platforms, i.e. the markets and the forums.

VII. Conclusion

We presented a descriptive analysis made on 4 main e-commerce platforms of the Darkweb: Silkroad 2, Agora, Alphabay and Nucleus. This analysis targeted the entire duration of these sites' activity, from 2013 to 2015. We focused our attention on collecting ads and services related to malicious hacking and cyber criminality. By studying the total value of the products offered on the different markets analyzed, the most striking result was that it is not a question of small-scale delinquency. We are talking about markets worth up to millions of USD each year. It is a scale of organized crime, well anchored in the field and whose demand and supply is booming. Also we were able to highlight the existence of well organized cells that monopolize the market. Indeed, 98% of the market is controlled by only 20% of sellers. In addition to this, it is a market that is based mainly on the reputation of sellers and the feedback of customers. Users are generally satisfied by the outcome of their transactions with cyber-criminals.

Acknowledgment

This work is partly funded by the NAT0 Science for Peace and Security (SPS) program under research contract SPS G5319 ”Threat Predict: From Global Social and Technical Big Data to Cyber Threat Forecast”[3].

References

[1] Benjamin, Victor, et al. ”Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops.” Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on. IEEE, 2015.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] Nunes, Eric, et al. ”Darknet and deepnet mining for proactive cybersecurity threat intelligence.” arXiv preprint arXiv:1607.08583 (2016).

[4] Greenberg, Andy. ”Drug market agora replaces the silk road as king of the dark net.” Wired. URL http://www. wired. com/2014/09/agora-bigger-than-silk-road (2014).

[5] Hosking, Jonathan RM, and James R. Wallis. ”Parameter and quantile estimation for the generalized Pareto distribution.” Technometrics 29.3 (1987): 339-349.

[6] C. Efroymson, ”The Kinked Oligopoly Curve Reconsidered,” The Quarterly Journal of Economics, vol. 69, no. 1, p. 119, 1995.

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] S. Chakrabarti, M. Van den Berg, and B. Dom. Focused crawling: a new approach to topic-specific web resource discovery. Computer Networks, 31(11):16231640, 1999.

[9] S. Chakrabarti, K. Punera, and M. Subramanyam. Accelerated focused crawling through online relevance feedback. In Proceedings of the 11th international conference on World Wide Web, pages 148159. ACM, 2002.

[10] T. Fu, A. Abbasi, and H. Chen. A focused crawler for dark web forums. Journal of the American Society for Information Science and Technology, 61(6):12131231, 2010.

[11] H. Chen. Dark web: Exploring and data mining the dark side of the web, volume 30. Springer Science and Business Media, 2011.

[12] Benjamin, Victor, et al. ”Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops.” Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on. IEEE, 2015.

[3] https://threatpredict.loria.fr