DARK WEB DATA ACQUISITION AND ANALYSIS USING SELF RELIANT CRAWLER

*

HIDAYAT-UR-REHMAN
Cyber Security department
Air University
Islamabad, Pakistan
211833@students.au.edu.pk

Abstract—**The dark web is a subset of the deep web, and it is defined as hidden websites that require special software to access. Users can use the dark web in the hopes of sharing files and information with minimal chance of discovery. Therefore this project will analyze the data on the dark web. It is important to have fresh data from the dark web that is why a semi-automated dark web crawler is required to collect data from data web different marketplaces like agora, silk road, etc. In this project, I have developed a crawler that uses queueing technique to collect URLs from the dark web and then the queue starts collecting data in textual form using the beautiful soup library, the data is stored in HTML & textual form, and stored in mongo database for future query application. The data is fed into an elastic search cluster to get a quick response result of any query from the frontend flask application. The analysis results show that the respective result like Drug markets, gun markets, red rooms, Data stealing, and hackers. The dark web is not always dark only 25% of it is dark and 75% is white which means the dark web is not only for illegal activities but there is also a good part of the dark web like the freedom to write, and freedom to publish any news without revealing identity.**

Index Terms—**Dark Web, Crawler**

I. INTRODUCTION

Internet, Media academics, and other sources of knowledge have several discussions regarding the Dark Web and Deep Web. Because there are no significant and official technical distinctions, the use of these phrases might become confusing. As a result, these phrases are commonly interchanged at different levels of hysteria. The following are the most often used and recognized definitions for both of these topics. The term ”Deep web” is to describe any type of material that conventional Internet/Surface web has not been kept up to date, for a variety of technical and non-technical reasons. The crawlers of conventional search engines like Google, Bing, Yahoo, etc. often conflict with the “Surface Web”, which is easily found and accessible by conventional search engines. Deep web content, for instance, can be password protected and required a special login, encrypted data Indexing cannot be authorized by the owner, or it cannot be easily hyperlinked due to complex hash URL form. Naturally, most of this material can be considered underground Activities, for example, many hacker forums are Accessible by password-protected techniques, Hackers use this technique for their ease to quickly access their content from anywhere in the world. The Deep Web, on the other hand, comprises many online sites and servers that serve Noble enterprises and information, including officially recognized web pages and databases that are hidden from public view. Hidden Facebook or Twitter accounts, according to experts, are indeed a portion of the Deep Web. The Dark Web, on the other hand, is a subset of the Deep Web that cannot be accessed with standard web browsers and instead requires specialist software. Access to private networks. As a result, explicit actions must be made to get access to The Dark Web, which is completely anonymous for both the user and the server. Many services, such as the Invisible Internet Project (I2P), Tor (the onion router), and pair-to-pair access, provide access to anonymous networks (P2P). However, the Tor Project’s hidden services remain the most common Component of the Darknet. Most journalists share their ideas, research, and reports on a dark web platform because truth costs a lot in our society, journalist they know if share the truth against any power, then we know the result, some unknown will kill them a very next day. if anybody shares something about the government and the government did not like it, in reaction he/she may go to jail. For freedom of speech, say anything at any time dark web is the platform that helps anybody to hide their identity from source to end user. From another perspective Yes dark web is the main source of crime because there is no worry about identity even transactions occurred in Bitcoins, the cryptocurrency which operates in the blockchain. If someone says that dark is fully Dark, according to my research, they are wrong because, on the dark web learning, research and academic work are more than bad content, in simple words, it is the technology like a blade either you must shave or you must cut the throats of others. The main purpose of this project is to analyze the dark web to reduce the fear and the answer the myths, about why people prefer the dark only for crime and why not explore the positive side of the dark web. The main purpose of this research is to analyze the various criminal and noncriminal activities on Dark Web like drug dealing, child abuse, hiring professional killers, Fake identity selling, credit card stealing, etc. The dark web is the main source of all untraceable criminal and noncriminal activities. Identification of those activities is the main aim of this research.

The core objectives of this research are as follows:

- To determine the beneficial and negative activities facilitated by the dark web.
- To determine the methods that reduce the dark web's harmful effects.
- To discover a more beneficial method of using the dark web.
- To find out why the dark web is the main source of cybercrime.
- To find which activity is on trending on the dark web.

II. Related work

In recent years, web forums both in and out of the Dark Web have been a hotbed of research, with authors tackling them from a range of perspectives, covering cyber security and intelligence. The works that are most similar to ours are on subsurface crawling mechanisms. Pastrana et al. [6] have just developed a technique that examines cybercrime outside of the Dark Web. The authors examine four English-speaking groups on the Clear Net and highlight the difficulty of browsing subterranean forums. Nunes et al. [15] collect threat intelligence from Dark Web and Deep Web sites and markets. Benjamin et al. [16] investigate cyberattacks on the dark web, focusing on stolen identity data activity as well as possible attack paths and software weaknesses. The writers collect information from carding businesses, IRC, and online forums, but they don't look into Tor Hidden Facilities.

The researchers of [17] and [18] examine prominent hacker groups in the United States and China, to identify key individuals, expertise, and connections in open online forums. They developed a system for extraction of features based on text mining and interaction integrity evaluation.

Motoyama et al. [19] do the same thing, looking at six distinct darknets on the open web and creating a measuring program based on the past. The comprehensive quantitative data analysis includes everything from top content to the extent of the contiguous client base to user activities and interactions.

Furthermore, their findings are based on stolen SQL dumps from the forums, whereas BlackWidow [22] is a platform that collects data in real time via the forums' interface.

We discover multiple business firms beyond the research literature that try to do a computerized evaluation of cyber security knowledge from the Dark Web and other channels. DarkOwl [20] and Historic Future [21], which analyze the Dark Web in many regions and claim to identify attacks, compromised data, and signs of penetration, are two examples.

III. Methodology

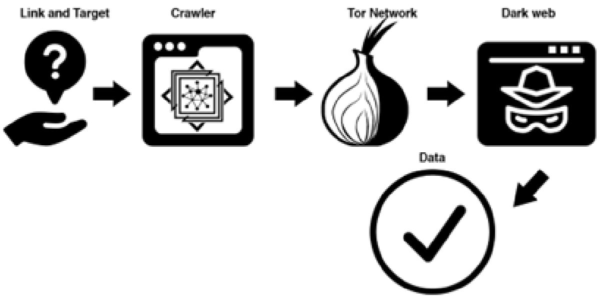

Fig. 1. Framework of the proposed model

The above methodology in Fig 1 shows that the user will take Dark web URL and Feed it to the crawler, Crawler will make sure that the URL is correct and exists in the domain of the dark web then the crawler will ask for specific fields which we want to crawl, Crawler will make sure the fields are present in that specific website and then the crawler will parse the Data present in fields and send it to the user.

The next part of the crawler will become into action and Data will be saved in HTML & Textual format. The Data pre-processing step will come into action and then the data will be fed into some statistical algorithm to generate a result and final the result will be displayed on screen in the form of graphs.

A. Tor Hidden Services

The Onion Router is the extended form of the word TOR. Tor enables us Anonymous to communicate through a secret relay network. We Apply Onion routing and binocular concepts, that the users gain anonymity by sending their request through a special network called circuits of onion network which consist of three or more relay nodes. Three is the minimum limit of relay nodes. Every tor circuit must contain at least three intermediate nodes.

The starting node through which our request enter is called the guard node and the exiting node through which our request leaves the circuit is called the exit node.

As tor is excellently a crowded network, these relays operate on a large-scale Volunteer network that has been voluntarily donated by people or organizations which is the most important tool for anonymous routing through the tor circuit. Through these Rely we get anonymity, even with banned Internet access some countries can access the restricted areas of the internet.

Tor is used because people want to be anonymous. Tor network is also used by intelligence secret services to run their relay nodes, through which they can communicate to reduce the risks to user's identity and ensure the reliability of data transmission. Regardless, Tor is the most popular network to hide Someone's identity on the Internet. In addition to enabling users to interact with the website anonymously.

Hidden services were introduced in 2004, providing service not only to the client but also to the server is also called the

anonymous respondent. Such use of Hidden services, any Internet service such as a normal web page, including forums or message boards in which we are interested to Hide their IP address from users.
When a client connects to the hidden service, all data transmission occurs through a point called Rendezvous point. This point is connected to anonymous Tor circuits which are connected to both client and Server. On the entire, there are five main parts of all hidden services.

1. self-hidden service
2. client side
3. Rendezvous Point
4. introduction point
5. Directory server

B. Working of Tor

When a user sends the request to the tor Client replay with a minimum of three nodes, every node has its own private and public key.
Every node in the Tor circuit responds with keys, thus user node keeps the private keys secret and establishes the connection with relay nodes in the TOR network using public keys when the user node sends data let's suppose D1, D1 data is encrypted by User node using all public keys of relays nodes which are received from Tor circuit during connection establishment, Every relay node has its key, therefore when relay node receives data D1, relay mode decrypt its headers using own private key and send it to next destination, this process continuous until exit node receives data D1. The exit node decrypts the data, reads the headers, and sends it to the destination server.

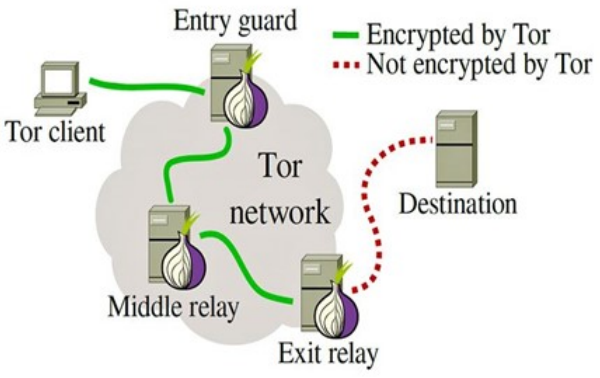

Fig. 2. Working of Tor Network

In response to data D1 from the server encrypt and send to the last node from which data D1 was received. All nodes which are way received a D1 response from the server use encryption until the guard node received a response of D1 guard node perform its encryption and send D1 to the users' node, users node has all the keys, User node decrypts the D1 one by one until readable data and display result on Tor browser. Fig 2 shows the working of the TOR network.

IV. Acquisition of Related Data

The challenging task for me is the identification of target data that is related to my research. Due to the secretive nature of the target data, there is no open-source data available that could be used as input to produce the result. According to my research 87% of Dark Web sites do not link to other onion websites, Dark Web is isolated short-life silos than the conventional Web (Surface web) which is more stable, reliable, and clear in structure.
There are also out-of-date gatherings of URLs (both surface web as well as Hidden Services) that exist on the Dark Web. Therefore, a fully automated method to overcome this issue is not suitable, that is why a semi manual method must be applied to collect data from the onion site or Dark web.

V. Parsing HTML data

Data retrieved over a crawler using URL-lib and beautiful soup, some part of HTML data is damaged or maybe coding mistake to overcome this type of data errors, LXML parser is used to parse the data with correct encoding sequence. Mining useful information from HTML data is very challenging depending on the complex HTML structure of the forums.
Extract data from HTML beautiful soup do a wonderful job. But every website has a different structure to extract useful data I must modify the crawler source code every time for every website. While this tactic may seem difficult, many HTML forums have a similar structure thus same code can be reused for different HTML forums.

The analysis is performed on the data that we collected using a crawler from the dark web. The main focus of this collection is the Agora marketplace. The dataset consists of 18 columns and 109692 rows. The columns include “Vendor, Category, Item, ItemDescription, Price, Origin, Destination, Rating, Remarks, BTC, Value, LogValue, Score, Deals, cat1, cat2, cat3, cat4” In the collected dataset we can explore:

- Countries of the Buyers and Sellers
- Distribution of Good Value
- Rating and Deal Count
- Vendors

VI. Results and discussions

In Fig 3 we can see most of the dealers are from the united states and the UK and so on. We can see the dealers from the United States lies in between 30000 and 35000 which is a very big number, then the most followed country is the United Kingdom which number lies in-between 10,000 and 15,000. Which is approximately 48% less than the United States dealer's number.

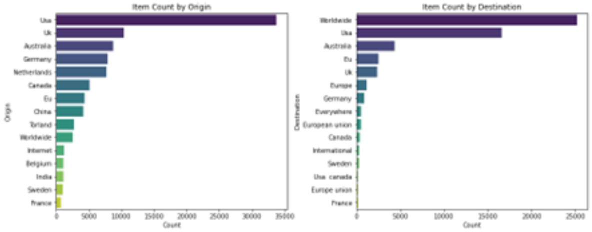

Fig. 3. Countries of the Buyers and Sellers

similarly, the lowest dealer on the dark web is from France which is in-between 0 to 5,000 which is 14% of unites states and 33% of the UK market dealers on the dark web. These are the discussion on dealers, and I think the same assumption can be taken place for buyers, but we cannot track the buyer's activity on the dark web because of anonymous behavior which is why we can assume the buyers is also from the same regions as sellers.

A. Log Dollar Value Distribution

In fig 4 analysis we are going to discuss the dollar value of items on the dark web, as we can see the dollar value starts from 101 and goes up to 108. moreover, the items price in dollars 104 is a very large number which means most of the items on the dark web can be purchased in-between 102-105, according to google the average income of the United States is 31,133 means 103, from which we can conclude that the dark web items are very cheap and average person can buy from the dark web.
These are just discussion not a claim that every citizen can a user of dark web.

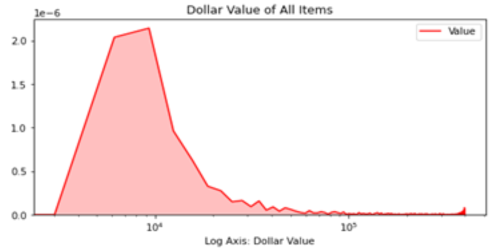

Fig. 4. Log Dollar Value Distribution

B. Rating and Deal Count

In Fig 5 rating and deal count on dark web agora marketplace are shown. As we can see that most of the seller's rating lies between 4 and 5. and a huge number lies between 5 it means that the dark web seller is reliable, and they can provide the best quality as they described in the form of the description which we can expand our understanding that dark web sellers of agora marketplace are trusted, and they provide the best of their services.

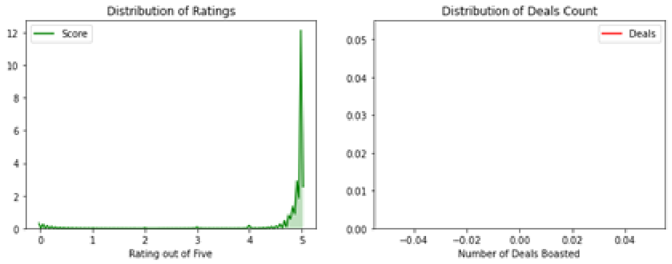

Fig. 5. Rating and Deal Count

C. Vendors Analysis

Illegal selling in dark is not a surprise in Fig 6 the number of vendors and the respective services is shown. we can see that optiman is the vendor who provides more than 800 services which can be various types of drugs etc.
If seeing the lowest vendor who provides around 200+ listings so this concludes that the dark web vendors provide services as many as they can. The active listing on the dark web lies in between 26 to 30 averages per person on the dark web.

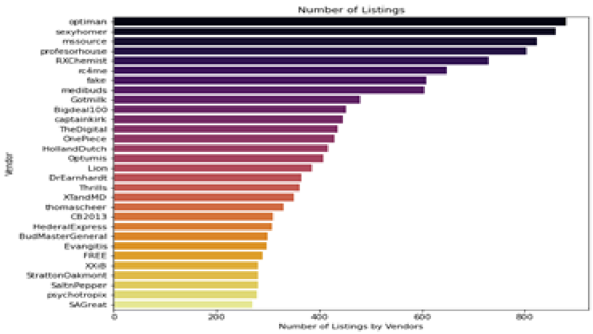

Fig. 6. Vedor Analysis

D. Dominant Category

Selling drugs on the dark web is not a surprising thing and its quantity is dominant all over the items is also not a new thing. In Fig 7 we can see that drugs are very dominant in the dark web agora marketplace.
The range of drugs on the dark web lies between 100-105 but the most reliable range is 102 and 103 and the outlier price range starts from 104 up to 105. if we conclude this discussion, we can see that in the agora dark web marketplace Drug is the most selling and heavy ordered good on the dark web.

E. Tier Two Category

It is interesting to note that while Cannabis has the most destructive and visible social consequences, seen in the current Cannabis overdose crisis, they are the most widely available on this Dark Web marketplace. If drugs can be further divided into sub-categories, then we get the results as shown in Fig 8,

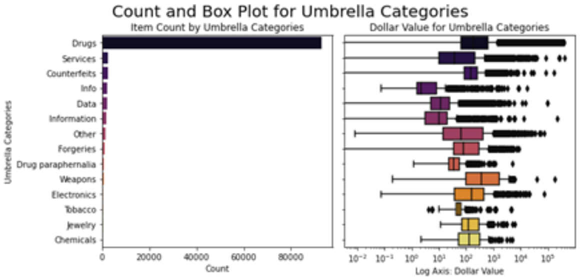

Fig. 7. Dominent Category

from the analysis of the dark web.

The subcategories of drugs on the dark web are ranked according to their available quantity. The most followed drug after cannabis is Ecstasy and stimulants which ranges from 10,000 and 15,000.

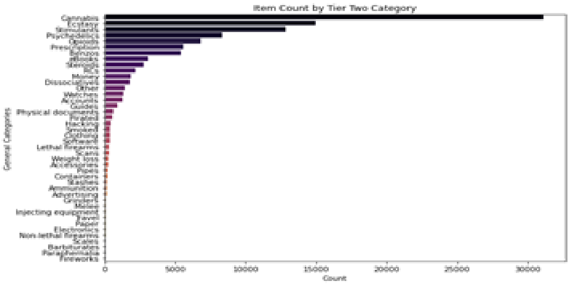

Fig. 8. Tier Two Category

F. Categories by Individual Umbrella Category

The second level category for services over the value of the dollar is shown in Fig 9. Money is the most highlighting factor in the graph which is 0-1400 and the hacking services lie between 0-450. The inter quartile range of services over dollar price lies between $100-1000 and the outlier expands the dollar value to $1000-10000.

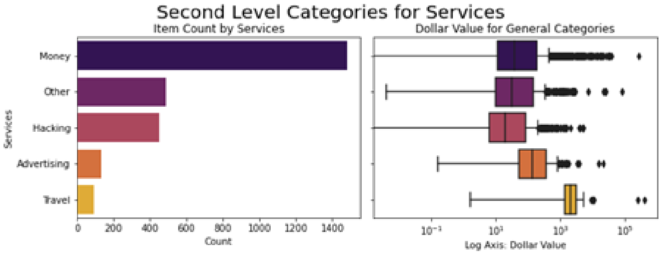

Fig. 9. Individual Umbrella Category

G. Forgeries

The analysis technique show in Fig 10 that the physical document forgeries is the most dominant and followed by scans which lies between 100-600 counts. The outliers lie between $1000-10000.

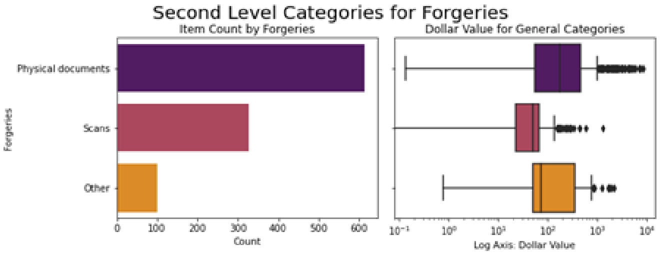

Fig. 10. Forgeries

H. Tobacco

The two main categories of tobacco are smoked and paraphernalia as shown in Fig 11. The smoked is the most selling on agora dark web marketplace.

The outlies of smoked and paraphenalia tobacco lies $10-100. The smoked tobacco is ranges form 0-400 count in agora marketplace of dark web.

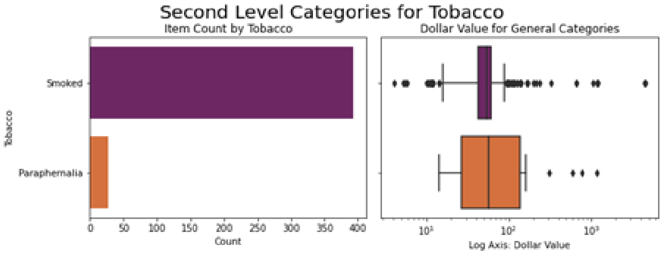

Fig. 11. Tobacco

I. Counterfeits

Analyzing the counterfeits from the agora marketplace of the dark web we found that the selling of steeled watches is in massive quantity as shown in Fig 12.

The PayPal and credit card are on the second number and clothing is followed by accessories and electronics. The range of counterfeits starts from 200-1200 and the detected outlier of counterfeits is 1000-10000.

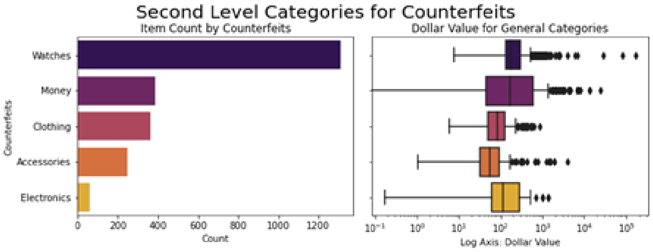

Fig. 12. Counterfeits

J. Data steeling

Data steeling is the main topic of discussion on the dark web. We see selling data on the dark web as usual things like someone is selling some product on Amazon or eBay. The most selling category of data is accounts, pirated software, and paid software as shown in Fig 13. in which dollar values lie between $100 and $1000.
Pirated software is a cracked version of paid software that can be embedded with malicious code through which a backdoor can be accessed by a hacker on the dark web. But this facility is aggressively available on the agora marketplace.

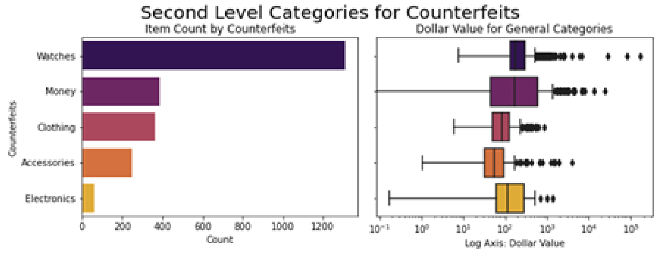

Fig. 13. Data steeling

K. Weapons

Lethal arms are those arms that are also called deadly weapons. Those arms use bullet types like AK-47s etc. during the analysis I come to know that the deadly weapons are more often on the dark web and their price is from $1000-10000 US dollars. The ammunitions are also the most fervent selling on the dark web and their price ranges from $100 to $400.
The price varies from weapon to weapon as shown in Fig 14, same as lethal weapons non-lethal weapons like katana, khanjar, etc. are also the discursive topic of dark web its price starts from 100 US dollar and ends on 200 US dollars. Overall, the discussion leads that illegal weapons are sold on the dark web in a very regular manner and the most interesting point is that they provide free home delivery all over the world.

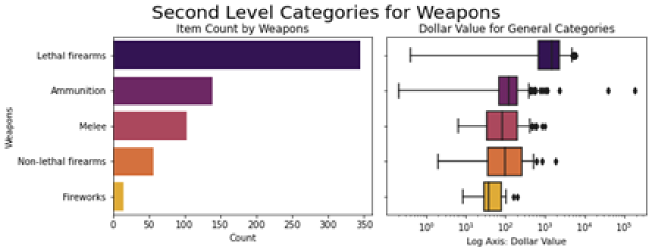

Fig. 14. Weapons

L. Three Tier Categories

The Fig 15 shows the overall three tier categories and its respective price range in the form of inter quartile range.

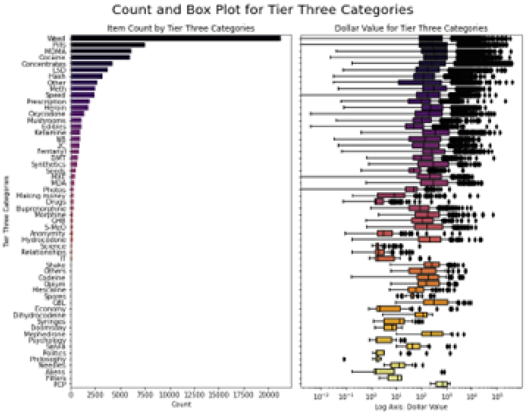

Fig. 15. Three Tier Categories

M. Cannabis

Agora marketplace is well known for drugs not only the quantity but also in quality and price. The agora market provides free home delivery for US and UK in drugs with escrow services and if the product is not delivered as it is mentioned in the description, they provide a money back grantee. Fig 15 shows the three-tier categories of drugs in which weed is the winner.
Weed's quantity ranges from 25,00-20,000 which is a very massive number, and the price range of weeds lies $10-1000 US dollars. Which count starts from 0-5000 approximately and its price range in US dollars is 100-10000 with the 105-outlier range. The mean prices of drugs on the agora marketplace are approximately between 100-500 us dollars. Last but not least Shake count is about 2000 and its price in US dollars means 500.

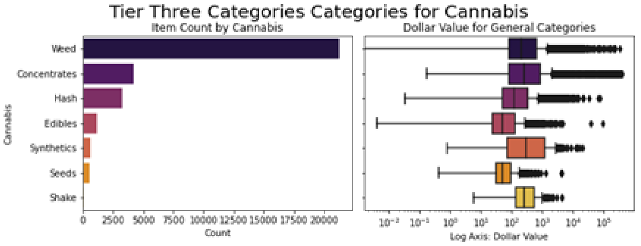

Fig. 16. Cannabis

N. Weapons Topics views

in Fig 17, the topic view of guns market in agora dark web, in which we can see that Dark web agora provides Escrow services. The highlight gun of the dark web is 9MM and Glock. Some other guns which are selling on the dark web are 9mm, luger, Glock new gen, stun gun, pistols, rifles, etc. each word in above words map shows its importance and its quantity and quality on agora dark web marketplace.

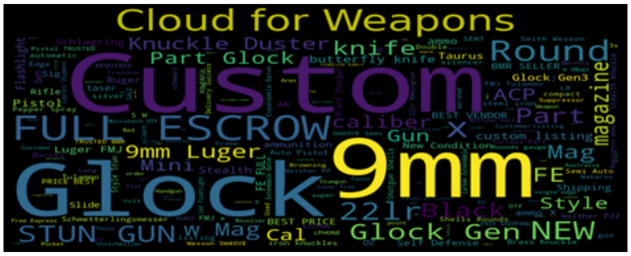

Fig. 17. Weapons Topics views

O. Data Topics views

Data stealing and providing stolen data is primary goal of dark web. The stolen data includes Porn portals, cloud data, emails, PayPal, Credit cards, any county national IDs, passports with lifetime access in just a few hundreds of US dollars as shown in Fig 18

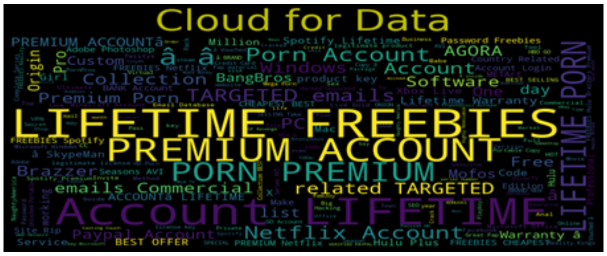

Fig. 18. Data Topics views

P. Drugs Topics views

Analysis discussion takes place in the above few paragraphs here I want to show the words map of drugs available on dark web agora marketplace. The most dominant word on the world map is Free shipping all over the world as shown in Fig 19, Full escrow is also the second dominant word in drugs.

XTC pill is also highlighted which means the word appears mostly in the dark web dataset.

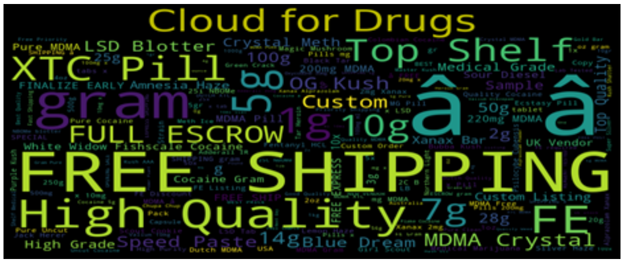

Fig. 19. Drugs Topics views

Last but not least this is my discussion on the dark web marketplace in which I discuss different variants of the dark web where we can find anything even, we can hire a professional assassin for just a few hundred US dollars to finish someone in the world. So, the conclusion dark web is a very dangerous place and the root of illegal activities.

I cannot say it is a totally bad place because on the dark web there are a lot of good things also like if someone writes a journal and he/she does not want to show identity they use the dark web to publish it. Some journalists also use it from reporting true news in which they feel their life will be in danger if they publish it with identity.

VII. Conclusion

In this research, I have collected approximately 16000 pages which consist of drugs marketplace, hitman, erotic pages, guns marketplace, and stolen information or data. By applying some analysis on agora market place we come to the result as expected. The result tells us about the various markets place for example drugs, stolen information, publish articles, and journals. It means the dark web is not only dark according to my research 25% is only dark and the rest is white. So we cannot say that the dark web is dark and according to the analyzing data it is clear there is some part of the dark web which is dark.

This research cover drug markets, weapons market, a sub category of drugs like opioids, cannabis, tobacco, etc. it also analyzes the stolen information and data like credit cards PayPal, etc., and also cover a small part of the erotic category and the results are discussed in the processing section.

References

[1] Intelliagg, "Deeplight: Shining a Light on the Dark Web. An Intelliagg Report," 2016.

[2] M. W. Al Nabki, E. Fidalgo, E. Alegre and I. de Paz, "Classifying illegal activities on TOR network based on web textual contents," in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017.

[3] A. Biryukov, I. Pustogarov and R.-P. Weinmann, "Trawling for tor Hidden Services: Detection, measurement, deanonymization," in IEEE Symposium on Security and Privacy (S&P), 2013.

[4] Hyperion Gray, "Dark Web Map," [Online]. Available: https://www.hyperiongray.com/dark-web-map/. [Accessed 7 1 2019].

[5] V. Griffith, Y. Xu and C. Ratti, "Graph Theoretic Properties of the Darkweb," arXiv preprint arXiv:1704.07525, 2017.

[6] S. Pastrana, D. R. Thomas, A. Hutchings and R. Clayton, "CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale," in Proceedings of the 2018 World Wide Web Conference, 2018.

[7] A. Pescape, A. Montieri, G. Aceto and D. Ciuonzo, "Anonymity Services Tor, I2P, JonDonym: Classifying in the Dark (Web)," IEEE Transactions on Dependable and Secure Computing, 2018.

[8] K. Bauer, D. McCoy, D. Grunwald, T. Kohno and D. Sicker, "Low-resource routing attacks against Tor," in Proceedings of the ACM Workshop on Privacy in Electronic Society, 2007.

[9] A. Biryukov, I. Pustogarov, F. Thill and R.-P. Weinmann, "Content and popularity analysis of Tor Hidden Services," in IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW) , 2014.

[10] I. Sanchez-Rola, D. Balzarotti and I. Santos, "The onions have eyes: A comprehensive structure and privacy analysis of Tor Hidden Services," in Proceedings of the 26th International Conference on the World Wide Web, 2017.

[11] L. K. Johnson, Ed., Handbook of intelligence studies, Routledge, 2007.

[12] "Puppeteer," [Online]. Available: https://pptr.dev. [Accessed 7 1 2019].

[13] Elastic,“Elasticsearch,”[Online].Available: https://www.elastic.co/products/elasticsearch. [Accessed 7 1 2019].

[14] P. Pons and M. Latapy, “Computing communities in large networks using random walks.,” Journal of Graph Algorithms and Applications, vol. 10, no. 2, pp. 191-218, 2006.

[15] E. Nunes, A. Diab, A. Gunn, E. Marin, V. Mishra, V. Paliath, J. Robertson, J. Shakarian, A. Thart and P. Shakarian, “Darknet and deepnet mining for proactive cybersecurity threat intelligence,” in IEEE Conference on Intelligence and Security Informatics (ISI), 2016.

[16] V. Benjamin, W. Li, T. Holt and H. Chen, “Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops,” in IEEE International Conference on Intelligence and Security Informatics (ISI), 2015.

[17] A. Abbasi, W. Li, V. Benjamin, S. Hu and H. Chen, “Descriptive analytics: Examining expert hackers in web forums,” in IEEE Joint Intelligence and Security Informatics Conference (JISIC), 2014.

[18] V. Benjamin and H. Chen, “Securing cyberspace: Identifying key actors in hacker communities,” in IEEE International Conference on Intelligence and Security Informatics (ISI), 2012.

[19] M. Motoyama, D. McCoy, K. Levchenko, S. Savage and G. M. Voelker, “An analysis of underground forums,” in Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, 2011.

[20] “DarkOwl,” [Online]. Available: https://www.darkowl.com. [Accessed 7 1 2019].

[21] “Recorded Future,”[Online].Available: https://www.recordedfuture.com. [Accessed 7 1 2019]

[22] Schäfer, M., Fuchs, M., Strohmeier, M., Engel, M., Liechti, M., & Lenders, V. (2019, May). BlackWidow: Monitoring the dark web for cyber security information. In 2019 11th International Conference on Cyber Conflict (CyCon) (Vol. 900, pp. 1-21). IEEE.