Content Analysis of Dark Net Academic Journals from 2010-2017 Using KH Coder

Prasanna Nattuthurai
Cal State LA
Pnattut@calstatela.edu

Arun Aryal, Ph.D.
Cal State LA
aaryal@calstatela.edu

Abstract

Networks that are not indexed by regular search engines such as Google, Bing, and Yahoo are called the darknet (Madden, 2015) and are available only to a closed group of people and accessible only via specific authorization, software, and configurations (Wood, 2010). The size of deep web content is about 400 to 550 times larger than the commonly defined world wide web (Beckett, 2009). Despite this vastness, Information Systems research lacks a general understanding of the nature and evolution of the darknet. To gain a deeper understanding and to classify the research, we analyzed over 391 abstracts, keywords, and titles of darknet related articles published in academic journals. Then we separated them into business and non-business disciplines. We performed Multi-Dimensional Scaling and Co-Occurrence Network analysis on the dataset using data analysis tool "KH Coder." Our results on these Business and Non-Business research papers led to the identification of major concerns on both domains. We suggest that further research should focus on positive aspects of the darknet as well.

1. INTRODUCTION:

Cybersecurity has been a concern for the business world and remains as a trending research topic. The importance of cybersecurity is illustrated by the statement of IBM CIO, " cybersecurity is the number one threat to every company in the world."[1] Furthermore, by the year 2015 cybercrime damage estimated to cost $3 trillion and it is expected to cost $6 trillion by 2021 all over the world (Morgan 2016). In research, cybersecurity is a cross-functional research domain for Information Systems, Business, Engineering and Computer Science.

The increase in cyber attacks and data breaches has raised a major concern for individuals to look for a more secure and anonymous way for data communication (Gregory C, 2015). One possible way to communicate anonymously is via darknet existing within deep web. While the terms darknet and deep web seem interchangeable, they refer to different elements of the web. Darknet is a part of the deep web that is approximately 400 to 550 times larger than the commonly defined world wide web (Beckett, 2009). It exists inside as layered proxy networks (Ali Rathore, 2016). VPN, proxy server and Tor network are one of the modern ways to stay anonymous. Tor network works on the principle of Onion Routing, in which messages are encapsulated in layers of encryption, analogous to layers of an onion. The figure 1 illustrates how onion router passes a message from one computer. The encrypted data transmits through a series of network nodes called onion routers, each of which "peels" away from a single layer, uncovering the data's next destination. When the final layer is decrypted, the message arrives at its destination. But the sender

[1] https://www.forbes.com/sites/stevemorgan/2015/11/24/ibms-ceo-on-hackers-cyber-crime-is-the-greatest-threat-to-every-company-in-the-world/#6f7313a73f07

remains anonymous since each intermediary knows only the location of the immediately preceding and following nodes (Goldschlag et al., 1999).

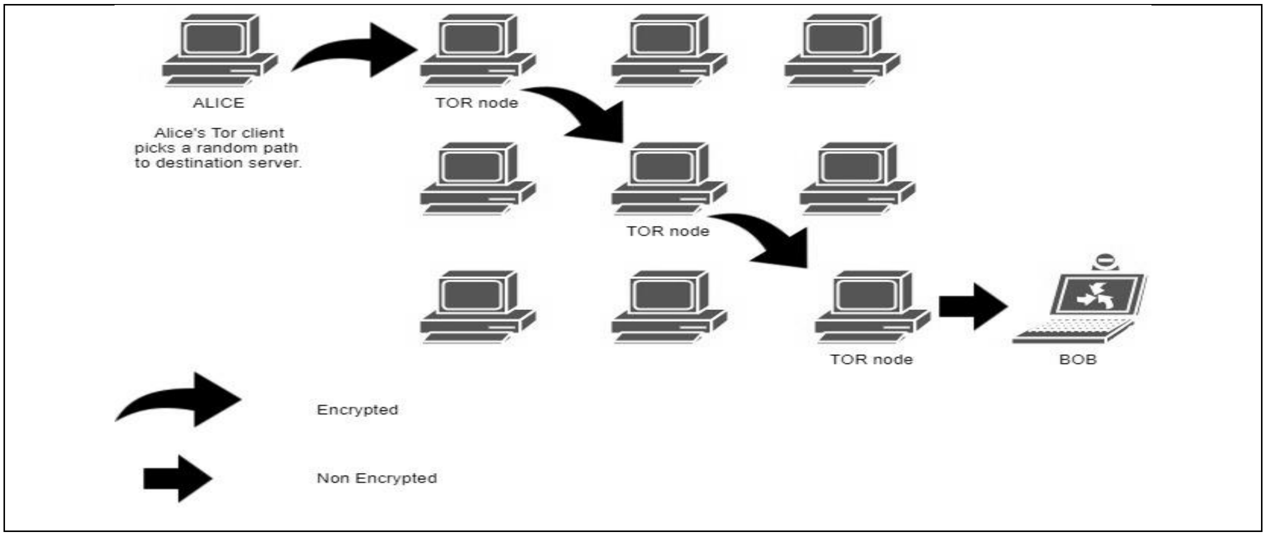
ALICE
Alice's Tor client picks a random path to destination server.
TOR node
TOR node
TOR node
BOB
Encrypted
Non Encrypted

Figure 1. Onion Routing

The United States Naval Research Laboratory mathematician Paul Syverson and computer scientists Michael G. Reed and David Goldschlag developed the core principle of Tor in mid-1990's with the purpose of protecting U.S. intelligence communications online. An agency within the department of defense, The Defense Advanced Research Projects Agency (DARPA) further developed onion routing in 1997 (Inc. The TorProject). Today, it is used every day for a wide variety of purposes by the military, journalists, law enforcement officers, activists, and many others. Tor Browser is an open source and free software created by Tor Project to access the darknet. I2P, Tails, and Freenet are the attractive alternatives to access darknet. Table 1 lists the topics discussed across the darknet into two classifications, the one being the number of sites in each category and the other represents the percentage of each classification. It shows that the Dark Net's content is diverse, with the largest number of sites being represented in the drugs category, but only by a small proportion (Owen, 2016).

Table 1. Darknet based Hidden Services metric in 2015 and 2016

2015		**2016**	
Category	**Percentage**	**Category**	**Percentage**
Guns	1.4	Violence	0.3
Chat	2.2	Arms	0.8
New (Not yet indexed)	2.2	Social	1.2
Abuse	2.2	Hacking	1.8
Books	2.5	Illegitimate pornography	2.3
Directory	2.5	Nexus	2.3
Blog	2.75	Extremism	2.7

Porn	2.75	Unknown	3
Hosting	3.5	Other illicit	3.8
Hacking	4.25	Finance	6.3
Search	4.25	Drugs	8.1
Anonymity	4.5	Other	19.6
Forum	4.75	None	47.7
Counterfeit	5.2		
Whistleblower	5.2		
Wiki	5.2		
Mail	5.7		
Bitcoin	6.2		
Fraud	9		
Market	9		
Drugs	15.4		

Moore and Rid also categorized the dark web based on a python web crawler methodology, “a script that cycled through known hidden services, found links to other dark websites, ripped their content, and then classified it into different categories.” In their methodology, if a page didn't display any content at all, or only had under 50 words, it was placed into the "none" category (Cox, 2016).

When each category is plotted against the percentage of Hidden Service (HS) requests it received we can see that a different picture emerges (see Figure 2). Requests to abuse sites represented more than 80 percent of total observed requests, although they accounted for only two percent of the total HS’s available (see Table 1). It is important to emphasize what is being measured (Owen, 2016).

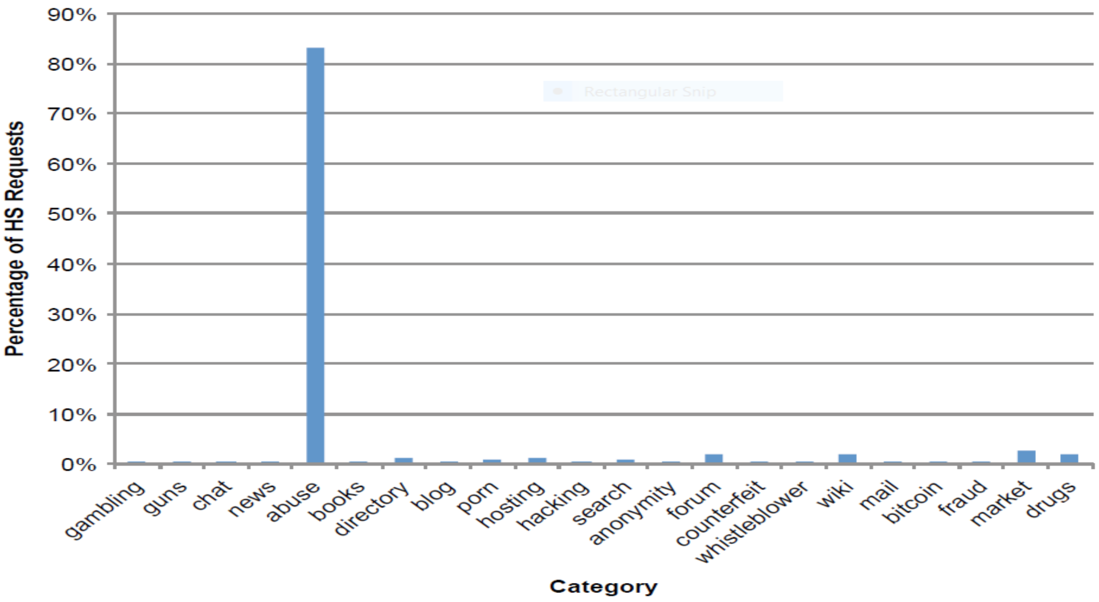

Figure 2. Percentage of Requests by Classification Category (Owen, 2016)

These HS are often difficult to read and can change periodically. One way to access them is to bookmark them or to check the HS websites like "Hidden Wiki" which serves as a directory for many other HS. For instance, the old Hidden Wiki URL http://kpvz7ki2v5agwt35.onion/ replaced by http://zqktlwi4fecvo6ri.onion/.

2. Data collections:

We gathered Darknet related academic journals from various databases such as IEEE, ABI, Engineering Village, ACM, etc. We applied advanced search method which performs a Boolean operation on the set of keywords given on the search engine. To narrow down the intended research, we chose Journal article format, English language, Information Systems and Computer Science as the subject area and darknet, dark web, tor etc as Keywords. Since darknet is relatively new, we limited the research published between 2010 – 2017. With the complete citation downloaded from the database, we created an organized data set using the Endnote software. Endnote process all the downloaded citations from different databases and created a standard data set. We exported the dataset provided by the endnote to an excel sheet; based on the area focused and the topic discussed we further categorized them into Business and Non-Business journals. Followed by that, we used Title, Abstract, Author, Year, Journal Source and Journal type as attributes. Since the data analysis tool, KH Coder couldn't process the data from excel as anticipated, we combined each abstract and its respective title into a paragraph in a notepad and then feed it for analysis.

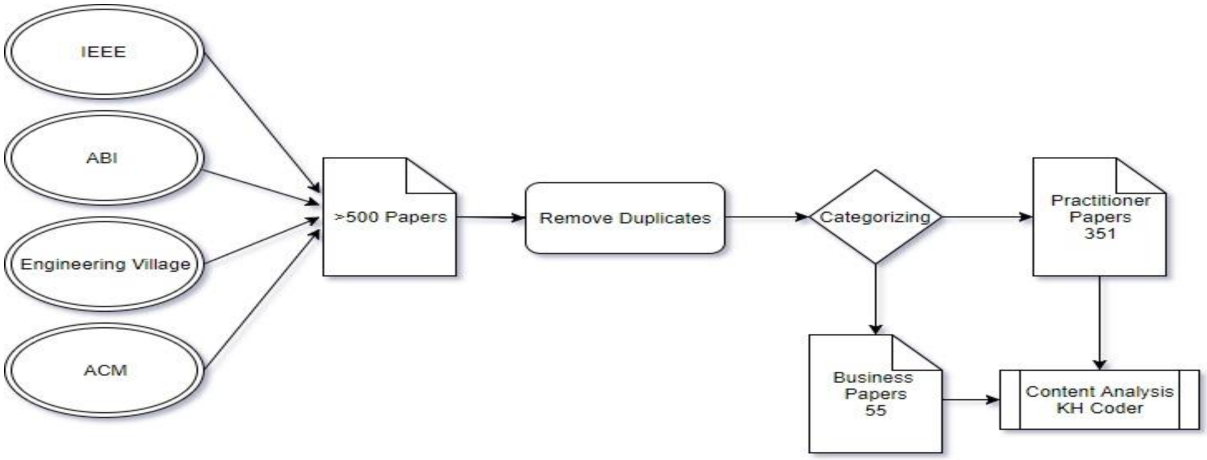

Figure 3. Methodology Flowchart [18]

3. Data Analysis and Results:

In this paper, we used KH Coder for data analysis. KH Coder is an open source software for quantitative content analysis, text mining and computational linguistics. By inputting raw texts, we can use searching and statistical analysis functionalities like KWIC (Key Word In Context), collocation statistics, co-occurrence networks, self-organizing map, multidimensional scaling, cluster analysis and correspondence analysis via back-end tools such as Stanford POS Tagger, Freeling, Snowball stemmer, MySQL and R. We imported the notepad created for both business and non-business into KH Coder for analysis. The file is then preprocessed using "Run Pre-Processing" command, executing this command segments sentences in a target file into words, and organizes these results as a database. With the processed data, we did Multi-Dimensional Scaling and Co-Occurrence Network of words (KH coder). By default, KH coder has a set of Stop Words to exclude from its analysis also we added few manually to eliminate them from analysis such as *abstract*, *(c)*, *IEEE*, *%*, *A*, *+*, *asphalt* etc.

3.1. Multi-Dimensional Scaling:

Multidimensional scaling (MDS) is a means of visualizing the level of similarity of individual cases of a dataset (Borg I & Groenen P, 2005). An example of classical multidimensional scaling applied to voting patterns in the United States House of Representatives. Where one red dot represents one member of the House from the Republican party, and each blue dot represents one member of the House from the Democrat party. It enables us to conduct analysis on the extracted words and to draw the results in 1- to 3-dimensional scatter diagrams. We can use these diagrams for finding combinations or groups of words that have similar appearance patterns (KH Coder, Scaling). This analysis uses the table generated with the [Export Document-Word Matrix] command with the variables indicating positions and document lengths removed. Among the three methods: Classical, Kruskal and Summon we choose Kruskal's algorithm as it is the most widely used among these three. Kruskal's algorithm is a type of minimum-spanning-tree algorithm it starts the tree from the cheapest edge by adding the next cheapest edge, provided that it doesn't create a cycle. For distance, we want Jaccard coefficient among Jaccard, Euclid, and Cosine, as its emphasis on whether specific words co-occur or not. Regardless whether the name appears once or ten times in a document, it is considered to "appear" and a word co-occurrence is calculated irrespective of appearance frequency, and it is useful for analyzing sparse data (H.C. Romesburg, 1984).

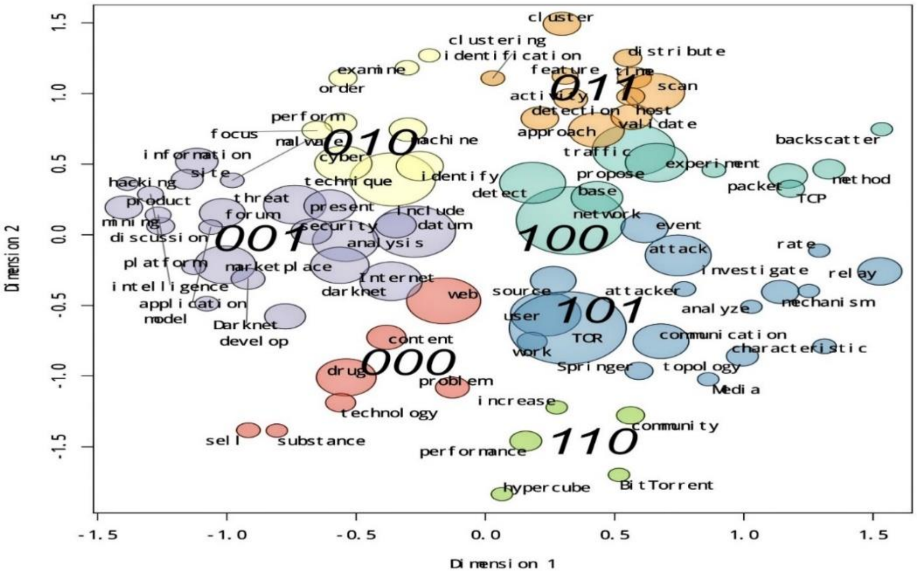

Figure 4. 2D scaling on Business Journals

In the 2D analysis, KH Coder performed a cluster analysis of words and indicate the result (clusters) by different colors. In Fig 4 Pink (000) cluster represents the theme “Drug Market”, Violet (001) cluster represents the theme “Security Threat”, Yellow (010) cluster represents the theme “Cyber Security Technique”, Orange (011) cluster represents the theme “Detection approach and Clustering”, Dark Green (100) cluster represents the theme “Network Signal

Traffic", Blue (101) cluster represents the theme "Tor Communication" and Green (110) cluster represents the theme "Bit torrent community".

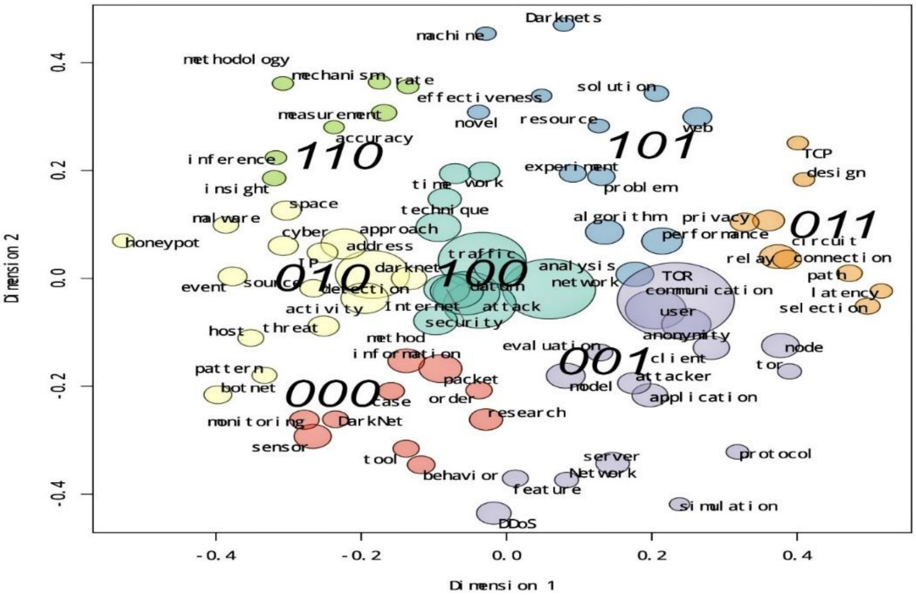

Figure 5. 2D scaling on Non-Business Journals

In Fig 5, Pink (000) cluster represents the theme "Packet Monitoring", Violet (001) cluster represents the theme "TOR Anonymity", Yellow (010) cluster represents the theme "IP address detection", Orange (011) cluster represents the theme "Circuit Connection", Dark Green (100) cluster represents the theme "Network Attack", Blue (101) cluster represents the theme "Algorithm Performance" and Green (110) cluster represents the theme "Accuracy Rate"

3.2. Co-Occurrence Network:

Co-Occurrence Network enables us to draw a network diagram that shows the words with similar appearance patterns, i.e. with high degrees of co-occurrence, connected through lines (edges) (KH Coder, Co-Occurrence Network). Since words connected with lines, it may be easier to understand the co-occurrence structures of the phrase, compared with the Multi-Dimensional Scaling, which only plots the words. It also shows the association between words and variables/headings, in addition to the connection between words.

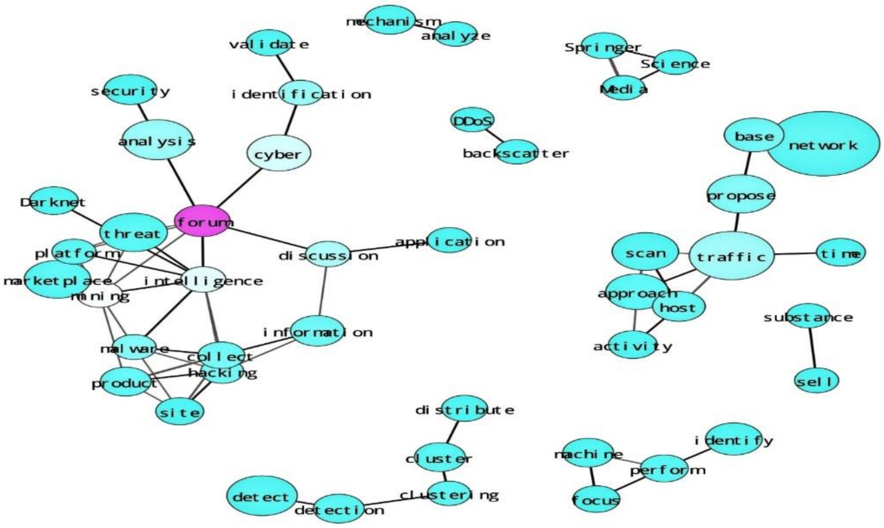

Figure 6. Co-Occurrence Network of words on Business Journals

In Fig 6, networks that are similar to the each of the clusters in Fig 5 are Forum-Marketplace-Product-Site, Security-Forum-Threat-Malware, Cyber-Forum-Intelligence, Cluster-Detect, Host-Traffic-Network, Mechanism-Analyze, and Focus-Perform-Identity.

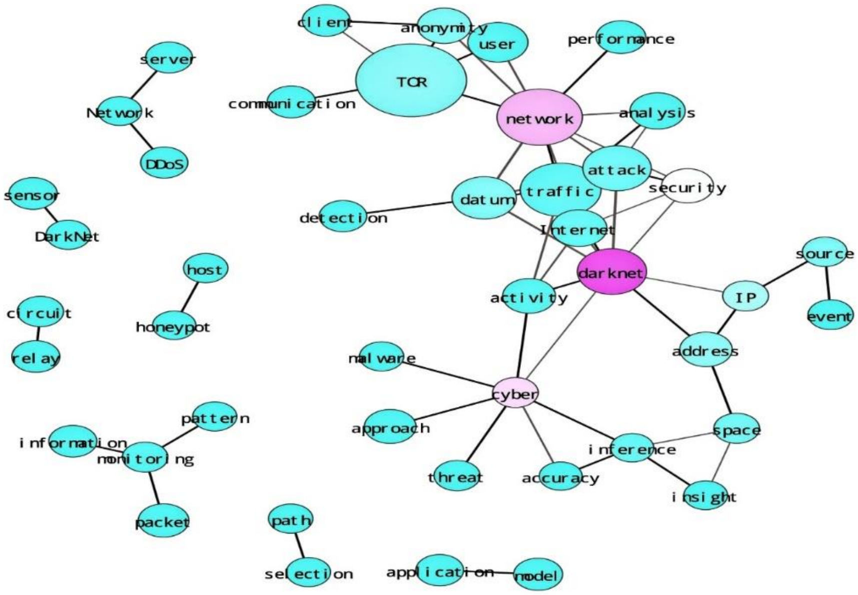

Figure 7. Co-Occurrence Network of words on Non-Business Journals

In Fig 7, networks that are similar to the each of the clusters in Fig6 are Packet-Monitoring-Pattern, Tor-Anonymity-Client, Event-Source-IP-Address, Circuit-Relay, Server-Network-DDos, Network-Performance and Accuracy-Interference-Insight

4. Concept Map:

From the results of Co-occurrence network and 2D scaling in Business and Non-Business, we identified the Networks that are formed relevant to the clusters created respectively. The Color code represents each cluster and Theme represents the relevant topic discusses in their respective

2D scaling output. Co-occurrence Colum in the table represents the Nodes that forms the network related to the respective themes.

Color	Theme	Co-Occurrence
Pink (000)	Drug Market	Forum-Marketplace-Product-Site
Violet (001)	Security Threat	Security-Forum-Threat-Malware
Yellow (010)	Cyber Security Technique	Cyber-Forum-Intelligence
Orange (011)	Detection approach and Clustering	Cluster-Detect
Dark Green (100)	Network Signal Traffic	Host-Traffic-Network
Blue (101)	Tor Communication	Mechanism-Analyze
Green (110)	Bit torrent community	Focus-Perform-Identity

Table 2. Keyword relation between 2D scaling and Co-Occurrence Network on Business Journals

Table 2 and Table 3 shows the relevant keyword comparison between 2D scaling and their respective co-occurrence network for Business and Non-Business journals respectively.

Color	Theme	Co-Occurrence
Pink (000)	Packet Monitoring	Packet-Monitoring-Pattern
Violet (001)	TOR Anonymity	Tor-Anonymity-Client
Yellow (010)	IP address detection	Event-Source-IP-Address
Orange (011)	Circuit Connection	Circuit-Relay
Dark Green (100)	Network Attack	Server-Network-DDos
Blue (101)	Algorithm Performance	Network-Performance
Green (110)	Accuracy Rate	Accuracy-Interference-Insight

Table 3. Keyword relation between 2D scaling and Co-Occurrence Network on Non-Business Journals

5. Key findings from Data Analysis:

We reviewed both Business and Non-Business data analysis output from KH Coder. It reveals the major concerns in each of them and some concerns in common as well. From the firm perspective keywords such as *drug, mining, illegal, forum, backscatter, private*, etc. and from non-business perspective keywords such as *anonymity, probe, botnet, honeypot, mobile network*, etc. has the high frequency of occurrence respectively. Commonly discussed words are *network, attack, malicious, threat, identity, security and data*.

Business	Non-Business	Common
Drug	Anonymity	Network
Mining	Probe	Attack
Illegal	Botnet	Malicious
Forum	Honeypot	Threat
Backscatter	Mobile	Identity
Private	Resource	Security
Social	Protocol	Data

Table 4. Keywords with a high frequency of occurrence

6. Negative > Positive:

It is evident from the analysis that, most of the keywords were pointing to the negative aspects of Darknet. But there is a lack of debate on the bright side such as SCADA systems data transfer, banking data transactions, the bitcoin blockchain, Internet of Things data, etc. Tor is being used by ordinary people to protect their privacy from corporations, to research sensitive topics, to skirt surveillance, to circumvent censorships, unscrupulous marketers, and identity thieves. Journalists and their audience use Tor for its privacy and security to report without borders. Activists & Whistleblowers uses Tor to stay anonymous from the government and organizations they raise their voice against. Even Militaries use Tor; field agents use Tot on their deployed location to mask their internet activity.

7. Looking at the Big Picture:

Darknet is not always about bad guys doing illegal activities anonymously. Tor shows that there is a lot of technologies and standards which we can incorporate into our daily life in future.

7.1 Bitcoin:

The infamous crypto-currency which accounts for the primary mode of payment in the darknet which circumvents the traditional problems in conventional currency system (Brito &Castillo, 2013). The system is peer-to-peer, and transactions take place between users directly, without an intermediary. Network nodes verify these operations and recorded in public distributed ledger called the blockchain. Bitcoin is also a decentralized payment system, and the transaction fee is optional; which avoids the major problem of the cost associated with resources, infrastructure, overseas trade, etc.

7.2 Message encryption:

Tor uses TLS (Transport Layer Security) protocol to establish the onion routing circuit. On top of this many Tor, sites use PGP (Pretty Good Privacy) encryption program to encrypt the IDEA symmetric key for the message using asymmetric encryption such as RSA/DSA/ElGamel. The asymmetric encryption is used only to initiate the connection between the website and user whereas symmetric encryption is used through the session as it is much faster.

7.3 Censorship:

Darknet is one of the easiest ways to achieve net neutrality. For instance, darknet websites can be accessed all over the world. Local restriction on a traditional webpage can’t apply for hidden service or contents.

Conclusion:

Our findings are limited by the number of journals we analyzed in this study. In future, we will enlarge our research by incorporating new research papers and conference papers in our analysis. From our data analysis, it is evident that the current discussion on darknet focuses more on the negative aspects of its use. Instead, we should look at the Tor model as a base of anonymity and security for future development in data communication as discussed in section 7. We must be resilient in our future development of internet security and privacy.

References

1. *Shahzeb Ali Rathore (7 November 2016). "Deep Web: The Dark Side of IS." RSIS Publications.*
2. *Beckett, Andy (26 November 2009). "The dark side of the internet." https://www.theguardian.com/technology/2009/nov/26/dark-side-internet-freenet*
3. *Borg, I.; Groenen, P. (2005). "Modern Multidimensional Scaling: theory and applications (2nd ed.)."*
4. *Jerry Brito & Andrea Castillo (2013). "Bitcoin: A Primer for Policymakers". Mercatus Center. George Mason University. Retrieved 22 October 2013.*
5. *Cox, Joseph (1 February 2016). "Study Claims Dark Web Sites Are Most Commonly Used for Crime."*
6. *Goldschlag D., Reed M., Syverson P. (1999.) Onion Routing for Anonymous and Private Internet Connections, Onion Router.*
7. *geography.oii.ox.ac.uk/wp-content/uploads/2014/06/Tor_Hexagons.png*
8. *Gregory C. Wilshusen., (2015) "Cyber Threats and Data Breaches Illustrate Need for Stronger Controls across Federal Agencies." United States Government Accountability Office*
9. *H.C. Romesburg (1984). "Cluster Analysis for Researchers. Belmont, Calif. Lifetime Learning Publications."*
10. *https://www.torproject.org/about/overview.html.en*
11. *http://sourceforge.net/projects/khc/*
12. *https://sourceforge.net/p/khc/wiki/Multi-Dimensional%20Scaling/*
13. *https://sourceforge.net/p/khc/wiki/Co-Occurrence%20Network/*
14. *Mary Madden., (2015) "Public Perceptions of Privacy and Security in the Post-Snowden Era." Pew Research Center.*
15. *Moore, Daniel. "Cryptopolitik and the Darknet." Survival: Global Politics and Strategy.*
16. *http://www.csoonline.com/article/3110467/security/cybercrime-damages-expected-to-cost-the-world-6-trillion-by-2021.html*
17. *Owen, Gareth. "Dr. Gareth Owen: Tor: Hidden Services and Deanonymisation."*
18. *Wood, Jessica (2010). "The Darknet: A Digital Copyright Revolution." Richmond Journal of Law and Technology.*
19. *Ylijokia, Ossi and Porrasa, Jari (2016). "Conceptualizing Big Data: Analysis of Case Studies"*

GitHub:

https://github.com/prasannavarshan/Dark-Net.git