Research Proposal Title

“An Empirical Study of Deep Web based on Graph Analysis”

Author: **Md Monzur Morshed**
m.monzur@gmail.com

DevSocial Research & Management Consulting
(www.devsocial.org)

Abstract

The internet can broadly be divided into three parts: surface, deep and dark among which the latter offers anonymity to its users and hosts [1]. Deep Web refers to an encrypted network that is not detected on search engine like Google etc. Users must use Tor to visit sites on the dark web [2]. Ninety six percent of the web is considered as deep web because it is hidden. It is like an iceberg, in that, people can just see a small portion above the surface, while the largest part is hidden under the sea [3, 4, and 5]. Basic methods of graph theory and data mining, that deals with social networks analysis can be comprehensively used to understand and learn Deep Web and detect cyber threats [6]. Since the internet is rapidly evolving and it is nearly impossible to censor the deep web, there is a need to develop standard mechanism and tools to monitor it. In this proposed study, our focus will be to develop standard research mechanism to understand the Deep Web which will support the researchers, academicians and law enforcement agencies to strengthen the social stability and ensure peace locally & globally.

Keywords: dark web, cybercrime, law enforcement

Introduction

The Dark Web, a conglomerate of services hidden from search engines and regular users, is used by cyber criminals to offer all kinds of illegal services and goods [35]. Cybercriminal activities in the dark web can be considered one of the critical problems for societies around the world [5]. Web mining techniques such as content analysis and structure analysis can be useful for detecting and avoiding terrorist's threats all over the world [7]. Nowadays social network analysis (SNA) is used to study a variety of economic and organizational phenomena and processes [6, 8, 9, and 10]. Social network analysis (SNA) is used effectively to counter money laundering, identity theft, online fraud, cyber-attacks, and others. In particular, the SNA methods are used in the investigation of many illegal operations with securities and investments, for the prevention of riots and others [6, 11, and 12].

Graph theory has long been a favored tool for analyzing social relationships [13, 14] as well as quantifying engineering properties such as search ability [13, 15]. For both reasons, there has been numerous graph-theoretic analysis of the World Wide Web (www) from the seminal [13, 16 - 20] to the modern [13, 21]. Graph theory as a tool can be used for analyzing social relationships for the dark web [13].

SNA [34] is a graph-based method for analyzing social relationships and their impact on individual behavior and organizational structure. It was developed by sociologists and has been applied in many academic fields such as epidemiology and Computer-mediated communication (CMC). After classifying and clustering the captured data, the characteristics of the special participants can be extracted. Through the social network analysis method, the social interaction mode with other cybercrime, the type of content published, and the frequency of discussion of the participating topics can be obtained [27].

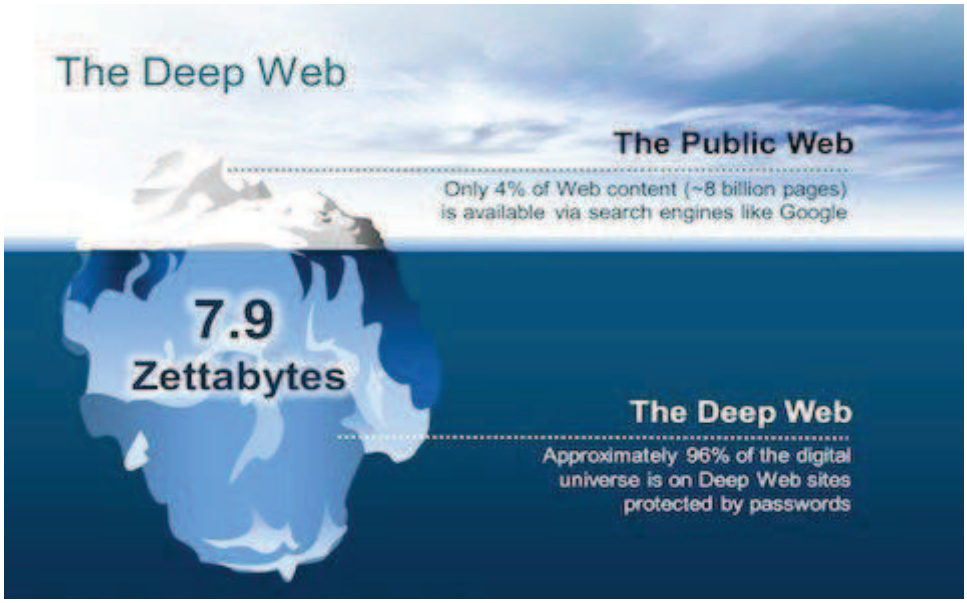


Figure 1: The deep web comparing with the visible web [22]

The dark web can provide anonymity by implementing of onion routers, which encrypt and bounce communication through a network of relays run by volunteers around the world [5]. The United States Naval Research Laboratory has developed the onion router (Tor) for anonymity to protect sensitive information and network. Tor program was released to the Internet users in 2004 [5]. It can provide privacy, and encryption, direct Internet traffic by using a series of virtual tunnels. It can help users to reach blocked contents and destinations. The Tor website ends with .onion, while other web domains end with .net, .com, .edu, .org, …etc, and can be opened by using the Tor software [5, 23]. Other programs can provide anonymity with encryption such as: ZeroNet, GNUNet, FAI (Free Anonymous Internet), and Freenet [5, 23].

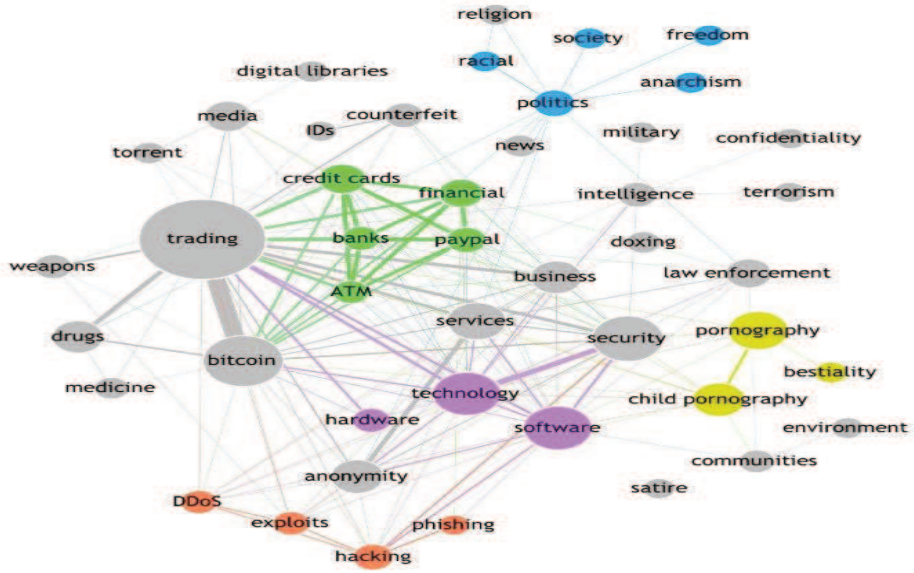


Figure 2: A topic taxonomy of Tor Hidden Services [24]

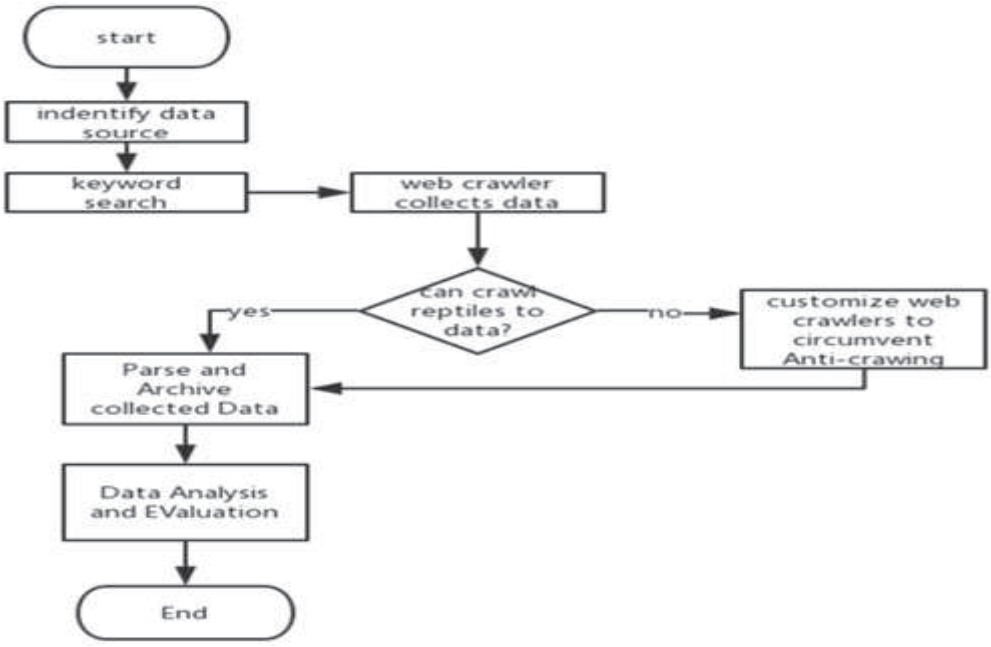


Figure 3: Dark Web crawler flow chart [27]

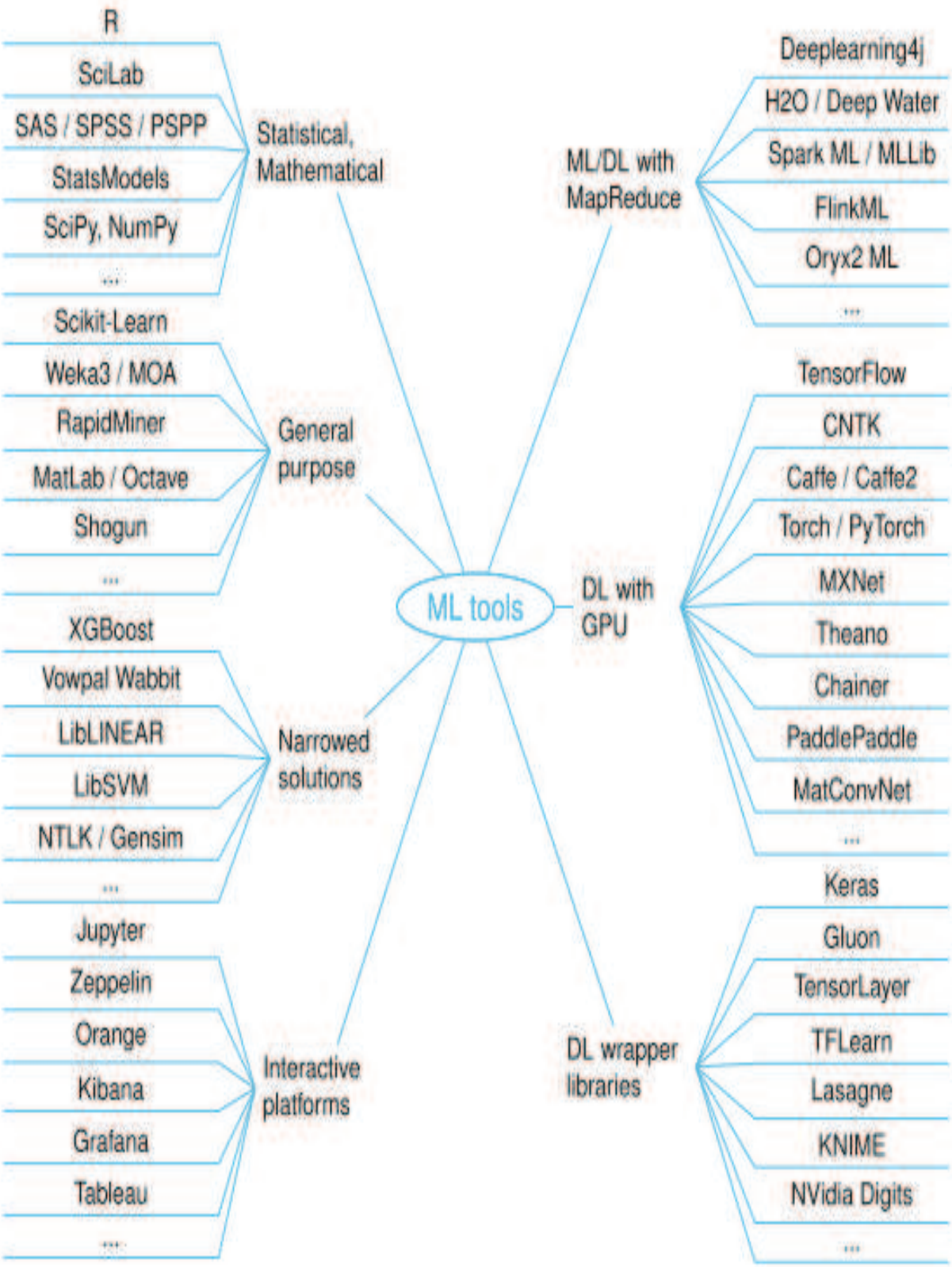
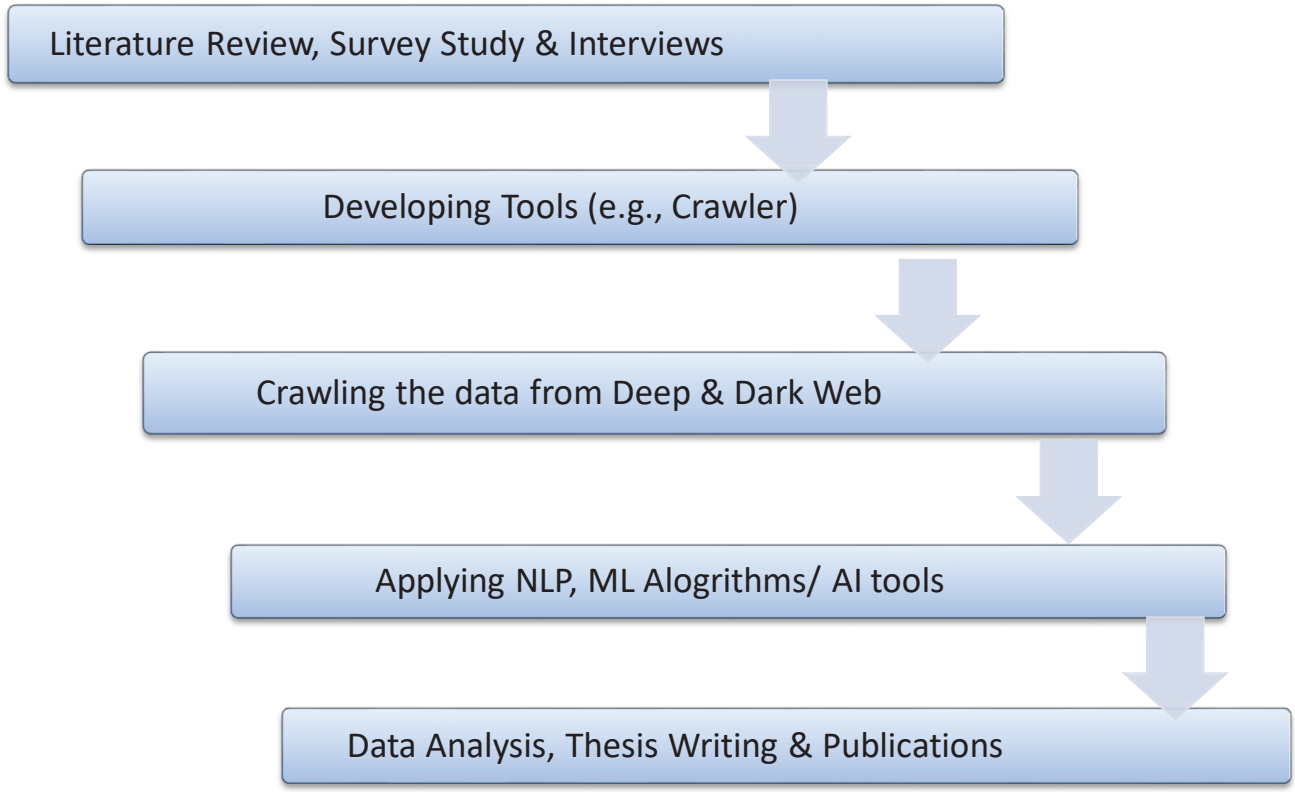


Figure 4: Overview of Machine Learning frameworks and libraries [38] (G. Nguyen et al)

Research Problem

One area that has not received adequate attention in the vast academic literature surrounding extremist movements and their use of the Internet is the Dark Web, whose websites are vaguely assumed to work as hubs for terrorists, drug-traffickers, and gangs [49]. With the rise of technology, cyber criminals are becoming more and more empowered. On the other hand, law enforcement agencies do not have adequate resources and technologies to fight cyber-crimes and monitoring activities on dark web. One of the primary challenges posed by the Dark Web to national security professionals is segregating out the "noise" from issues of legitimate national security concern. With annual cybercrime revenue estimated at approximately $1.5 trillion and considering the existence of 7,000-30,000 TOR sites, knowing where to look requires us to bound our focus to specific subject areas [50]. To find latest researches on dark web, an online query was executed on IEEE Xplore. The query result showed that only 250 resources available on dark web. Among those, 111 conference papers, 96 journals articles, 23 magazines, 13 books and 7 other resources. Therefore, we can say that deep web needs more academic attention to fight cyber-crime in this information age. By conducting research on "Deep Web", our primary focus is to deliver a comprehensive road map to fight cyber-crimes and devise new strategies to monitor deep & dark web while developing standard software systems.

Research Methodologies



Related Research Study

Latent Dirichlet Analysis (LDA) technique has been applied by [25] to discover latent topics in dark Web page's contents. LDA is a generative model to detect topics in a text corpus by determining likelihoods of each document, and then capture word and documents that being capable of exchange. Finding the threaten topics can assets detecting community key-members. A work done by [26] to extracting group key members using LDA to find the terrorize topics by integrating the LDA in dark Web portals to enhance the Social Network Analysis (SNA). Using the method can help to measure the radical of the member and assort the kind of member to expert or key-based on the selected topic. This work limited to dark websites use English language as communication language and it also done based in only one forum.

Zhang Xuan, a member of the Shandong Police College, and the Secretary of the Department of Information Security and Cryptography (CISC) of the University of Hong Kong, Professor Jinpei ou, co-published the Dark Net Threat Intelligence Analysis Framework, which proposes a concept of a hidden threat intelligence analysis framework. To help analyze crime traces in the dark network [27, 28].

In a recent work [29, 32], Qin et al. performed an empirical study of different global extremist organizations on the Web and presented how sophisticatedly they propagate their ideologies. Several studies have focused on sentiment analysis, opinion mining and affect analysis of user posts in Web forums [30], and the discovery of user roles and their ties have been appraised [31].

In a research study, Yang et al. [33] came up with a spectral coherence based clustering approach to identify dark Web clusters, which considers the temporal coherence of user activeness rather than contents or links as the primary information. They represented a group of users as a m-dimensional multivariate process which is used to derive the spectral density matrix and finally spectral coherence score is computed to identify the clusters [32].

Pastrana et al. [36] recently built a system that looks at cyber-crime outside the Dark Web. The authors discuss challenges in crawling underground forums and analyze four English-speaking communities on the Surface Web. In contrast, Nunes et al. [37] mine Dark Web and Deep Web forums and marketplaces for cyber threat intelligence. They show that it is possible to detect zero-day exploits, map user/vendor relationships and conduct topic classification on English-language forums, results that we have been able to reproduce with BlackWidow [35].

Al-Nabki et al. [40] presented a web-text-content-based classification pipeline containing TOR dark net illegal activities. They have used two well-known text representation techniques (Frequency Inverse Document Frequency and Bag-of-Words) together with three different supervised classifiers (Logistic Regression, SVM, and Naive Bayes). With the help of Uniform Resource Locators (URL), Kan et al. [41] classified the web pages by extracting features where a URL is segmented into tokens using information-theoretic measures. Noor et al. [42] proposed an automatic deep web classification technique, named “Query Probing”, where they extracted

the content from deep web data sources. Besides, it is commonly used for supervised learning algorithms and “Visible Form Features” [39].

Nunes et al. [43] discovered 16 zero-day exploits by monitoring forum posts in Darknet marketplaces. To reduce training data labeling requirements, their binomial classification method combined supervised with semi-supervised classifiers (eg: Label Propagation and Co-Training). Unsupervised k-means clustering was applied to character level n-gram features in [44] and partitioned Dark Web marketplace products into 34 clusters.

Thomas et al. analyzed the way of cybercriminals’ communications and what they exchange in forums [45]. Pastrana et al. focused on finding cybercrime actors in a large underground forum [46]. For evaluating private interactions, Overdorf et al. developed a method for automatically labelling threads that are likely to trigger private messages [47]. These studies were used to explore the market of underground forums and the social relationships of members.

Masashi et al. [48] conducted a study to efficiently extract threat intelligence from the dark web by using machine learning as an “active defense” against cyber-attacks. Furthermore, focusing on the current situation that myriad forums are rampant on the dark web, they proposed a method to identify the characteristics of these forums. The experiment showed that "doc2vec", a neural network based tool, has high performance as a method of natural language processing and feature extraction in machine learning. MLP indicated high classification performance of 90% or more based on the number of datasets used in the experiment. This proved that the vectorization of doc2vec accurately represents the features of the posts. Furthermore, their experiment has shown that it is effective to use machine learning for posts on the dark web [48].

Timeline

- Month 1-5: Background reading, Literature Review and Group planning
- Month 6-12: Prototyping of Crawler, interface design
- Month 13-18: Development of the actual system based on the prototype
- Month 19-20: Overall integration
- Month 21-24: Testing, revision and integration
- Month 25-26: Documentation
- Month 27-30: R&D on NLP, ML, AI implementation
- Month 31-32: Data Analysis
- Month 32-34: Thesis Writing
- Month 34-35: Thesis Review& Feedback adjustment.
- Month 35-36: Final preparation for publication

Deliverables

The target end results of this effort are:

- Crawler and AI System's source codes
- Data Source
- Thesis/ Report
- Project Documentation
- Standard Operation Plan
- Test Report
- Any artifacts developed during research

Future Research Scope

The future research scope on Deep Web has enormous potential. With the rise of "Artificial Intelligence" and Web Technology, research on Deep Web would be the next game changer. Furthermore, this area of research involves emerging fields such as big data and advance intelligent computing, etc. More precisely, this will play a significant role in global eGovernance.

References

[1] Abhineet Gupta, The dark web as a phenomenon: a review and research agenda, 2018

[2] W. Park, "A Study on Analytical Visualization of Deep Web," *2020 22nd International Conference on Advanced Communication Technology (ICACT)*, Phoenix Park, PyeongChang,, Korea (South), 2020, pp. 81-83, doi: 10.23919/ICACT48636.2020.9061283.

[3] Deep Web Sites 2018 | Dark Web | Deep Web Links | Hidden Wiki , https://www.deepweb-sites.com/

[4] Walsh, D., A Beginner's Guide to Exploring the Darknet, https://turbofuture.com/internet/A-Beginners-Guide-to-Exploring-theDarknet, May 2018

[5] H. Alnabulsi and R. Islam, "Identification of Illegal Forum Activities Inside the Dark Net," *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, Sydney, Australia, 2018, pp. 22-29, doi: 10.1109/iCMLDE.2018.00015.

[6] Kirichenko Lyudmyla, Radivilova Tamara and Carlsson Anders, Detecting cyber threats through social network analysis: short survey, arXiv, 2018

[7] H. M. Alghamdi and A. Selamat, "Topic detections in Arabic Dark websites using improved Vector Space Model," *2012 4th Conference on Data Mining and Optimization (DMO)*, Langkawi, 2012, pp. 6-12, doi: 10.1109/DMO.2012.6329790.

[8] Easley, D., Kleinberg, J. (2010). Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press, 819 p.

[9] Jackson, Matthew O. (2010). An Overview of Social Networks and Economic Applications. Handbook of Social Economics. Retrieved from https://web.stanford.edu/~jacksonm/socialnetecon-chapter.pdf. Accessed 07 March 2017.
[10] Matthew A. Russell (2011). Mining the Social Web: Analyzing Data from Facebook,Twitter, LinkedIn, and Other Social Media Sites. O'Reilly, 332 p.
[11] Carley, K., Lee, J., Krackhardt, D. (2002). Destabilizing networks. Connections, 24(3), 79-92.
[12] Stohl, C., Stohl, M. (2007). Networks of Terror: Theoretical Assumptions and Pragmatic Consequences. Communication Theory, 17, 93-124.
[13] Virgil Griffith, Yang Xu and Carlo Ratti, Graph Theoretic Properties of the Darkweb, arXiv, 2017
[14] Borgatti SP, Jones C, Everett MG (1998) Network measures of social capital. Connections 21: 27–36.
[15] Wu P, Wen JR, Liu H, Ma WY (2006) Query selection techniques for efficient crawling of structuredweb sources. In: Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conferenceon. IEEE, pp. 47–47.
[16] Barab´asi AL, Albert R (1999) Emergence of scaling in random networks. science 286: 509–512.
[17] Barab´asi AL, Albert R, Jeong H (2000) Scale-free characteristics of random networks: the topology ofthe world-wide web. Physica A: Statistical Mechanics and its Applications 281: 69–77.
[18] Adamic LA, Huberman BA (2000) Power-law distribution of the world wide web. Science 287: 2115–2115.
[19] Kumar R, Raghavan P, Rajagopalan S, Sivakumar D, Tompkins A, et al. (2000) The web as a graph. In:Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of databasesystems. ACM, pp. 1–10.
[20] A.-L. Barabási (2001) The physics of the web.
PhysicsWeb http://wwwphysicsweborg/article/world/14/7/09
[21] Meusel R, Vigna S, Lehmberg O, Bizer C (2015) The graph structure in the web–analyzed on different aggregation levels. The Journal of Web Science 1.
[22] Onur Catakoglu, Marco Balduzzi, Davide Balzarotti, "Attacks Landscape in the Dark Side of the Web", ACM, 2017.
[23] Lewis, M., What Is the Dark Web – Who Uses It, Dangers & Precautions to Take, May 2018, https://www.moneycrashers.com/darkweb/
[24] Spitters, M., Verbruggen, S., and van Staalduinen, M. 2014. "Towards a Comprehensive Insight into the Thematic Organization of the Tor Hidden Services," Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint: IEEE, pp. 220-223
[25] L. Yang, F. Liu, J. M. Kizza, and R. K. Ege, "Discovering Topics from Dark Websites," in 2009 IEEE Symposium on Computational Intelligence in Cyber Security, 2009, pp. 175-179.
[26] G. L'Huillier, S. A. Rios, H. Alvarez, and F. Aguilera, "Topic-Based Social Network Analysis for Virtual Communities of Interests in the Dark Web," in ACM SIGKDD Workshop on Intelligence and Security Informatics, 2010, p. 9.
[27] Y. Yang *et al*., "Hadoop-based Dark Web Threat Intelligence Analysis Framework," *2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, Chongqing, China, 2019, pp. 1088-1091, doi: 10.1109/IMCEC46724.2019.8984106.
[28] Xuan Zhang ,"A Framework for Dark Web Threat Intelligence Analysis". International Journal of Digital Crime and Forensics, 2018, 10 (4)
[29] J. Qin, Y. Zhou, and H. Chen, "A multi-region empirical study on the internet presence of global extremist organizations," Information Systems Frontiers, vol. 13, no. 1, pp. 75–88, Mar 2011.
[30] A. Abbasi, H. Chen, S. Thoms, and T. Fu, "Affect analysis of web forums and blogs using correlation ensembles," IEEE Trans. on Knowl. and Data Eng., vol. 20, no. 9, pp. 1168–1180, Sep 2008.
[31] C. C. Yang, X. Tang, and B. M. Thuraisingham, "An analysis of user influence ranking algorithms on dark web forums," in ACM SIGKDD Workshop on Intelligence and Security Informatics, ser. ISI-KDD '10. New York, NY, USA: ACM, 2010, pp. 10:1–10:7.
[32] T. Anwar and M. Abulaish, "Identifying cliques in dark web forums - An agglomerative clustering approach," *2012 IEEE International Conference on Intelligence and Security Informatics*, Arlington, VA, 2012, pp. 171-173, doi: 10.1109/ISI.2012.6284289.
[33] C. C. Yang, X. Tang, and X. Gong, "Identifying dark web clusters with temporal coherence analysis," in Proceedings of the 2011 IEEE international conference on Intelligence and security informatics, ser. ISI'11. IEEE Press, 2011, pp. 167–172.
[34] Li xiaolei, "Design and Implementation of Platform for Social Network Analysis Based on Hadoop." School of Electronic and Information, Ningbo University, Ningbo 315175, China

[35] M. Schäfer, M. Fuchs, M. Strohmeier, M. Engel, M. Liechti and V. Lenders, "BlackWidow: Monitoring the Dark Web for Cyber Security Information," *2019 11th International Conference on Cyber Conflict (CyCon)*, Tallinn, Estonia, 2019, pp. 1-21, doi: 10.23919/CYCON.2019.8756845.
[36] 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.
[37] 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.
[38] Nguyen, G., Dlugolinsky, S., Bobák, M. et al. Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artif Intell Rev 52, 77–124 (2019).
[39] R. Biswas, E. Fidalgo and E. Alegre, "Recognition of service domains on TOR dark net using perceptual hashing and image classification techniques," *8th International Conference on Imaging for Crime Detection and Prevention (ICDP 2017)*, Madrid, 2017, pp. 7-12, doi: 10.1049/ic.2017.0041.
[40] Md Wesam Al Nabki, Eduardo Fidalgo, Enrique Alegre, and Ivan de Paz." Classifying Illegal Activities on Tor Network Based on Web Textual Contents", Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 1, pp. 35-43, (2017).
[41] Min-Yen Kan and Hoang Oanh Nguyen Thi.."Fast webpage classification using url features", In Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 325–326. (2005).
[42] Umara Noor, Zahid Rashid, and Azhar Rauf. "A survey of automatic deep web classification techniques", International Journal of Computer Applications, pp. 43–50, (2011).
[43] 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 2016 IEEE Conference on Intelligence and Security Informatics (ISI). Institute of Electrical and Electronics Engineers (IEEE), Sept 2016, pp. 7–12.
[44] E. Marin, A. Diab, and P. Shakarian, "Product offerings in malicious hacker markets," in 2016 IEEE Conference on Intelligence and Security Informatics (ISI). Institute of Electrical and Electronics Engineers (IEEE), Sept 2016, pp.187–189.
[45] K. Thomas et al., ''Framing dependencies introduced by underground commoditization,'' in Proc. Workshop Econ. Inf. Secur., 2015
[46] S. Pastrana, A. Hutchings, A. Caines, and P. Buttery, ''Characterizing eve: Analysing cybercrime actors in a large underground forum,'' in Proc. Int. Symp. Res. Attacks, Intrusions, Defenses. Cham, Switzerland: Springer, 2018, pp. 207–227
[47] R. Overdorf, C. Troncoso, R. Greenstadt, and D. McCoy. (2018). ''Under the underground: Predicting private interactions in underground forums.'' [Online]. Available: https://arxiv.org/abs/1805.04494
[48] M. KADOGUCHI, S. HAYASHI, M. HASHIMOTO and A. OTSUKA, "Exploring the Dark Web for Cyber Threat Intelligence using Machine Leaning," 2019 IEEE International Conference on Intelligence and Security Informatics (ISI), Shenzhen, China, 2019, pp. 200-202, doi: 10.1109/ISI.2019.8823360.
[49] Ghadah Alrasheed and Brandon Rigato, Exploring the Dark Web: Where Terrorists Hide? February 5, 2019. Online Source: https://carleton.ca/align/2019/illuminate-exploring-the-dark-web-where-terrorists-hide/
[50] Jason Rivera and Wanda Archy, The Role of the Dark Web in Future Cyber Wars to Come. February 21, 2019 Online Source: https://smallwarsjournal.com/jrnl/art/role-dark-web-future-cyber-wars-come