On Understanding the Dark Web through Graph Analytics

Kleanti Bashalli[1], Alexandros Karakasidis[1], Sophia Karagiorgou[2], and George Pantelis[2]

[1] Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece
{ics20020,a.karakasidis}@uom.edu.gr
[2] Ubitech Ltd., Chalandri, Athens, Greece
{skaragiorgou,gpantelis}@ubitech.eu

Abstract. The Dark Web, a protected digital graph, contains a wealth of hidden information that evades traditional search engines. This work delves into the mysterious landscape of the Dark Web by analyzing its structure using graph data from a Dark Web crawler. Using exploratory data analysis techniques, we uncover structure, trends, and insights that contribute to a deeper understanding of the content hidden in this dark digital world. This research is a critical step toward understanding the contents of the Dark Web and improving our ability to navigate the complexities of the digital underworld.

Keywords: Dark Web · Data Analytics · Web Mining · Web Crawling.

1 Introduction

The Dark Web is a part of the internet that goes beyond what traditional search engines can find. It is a realm where anonymity is the key that allows users to access websites and services that escape the attendance of the average user. Exploratory analysis within the web involves diving into its hidden facets to understand its structure, content and related pertaining activities. This type of analysis through the application of algorithms and visualization offers insights into a range of activities occurring within the Dark Web, ranging from cybercrime to underground markets and more.

The use of exploratory graph analytics is necessary to strengthen digital security and reduce potential threats. A key business case is to proactively detect emerging cyberthreats by monitoring illegal forums, marketplaces and communication channels on the Dark Web. By exploring these hidden corners of the Internet, cybersecurity experts can uncover conversations about new attack vectors, vulnerabilities and malicious exchanges. Furthermore, breaches and misuse of business data are some cases of incidents that might be identified by mining the Dark Web.

This information allows companies to not only strengthen their defenses against impending cyber threats, but also take proactive measures such as patching vulnerabilities and improving security protocols. Dark web forensics is therefore becoming a strategic imperative in the ongoing fight against cybercrime and

represents a valuable tool for staying one step ahead of malicious actors in an ever-evolving digital security landscape.

The main goal of this work is to systematically analyze the Dark Web through a dataset [8] gathered through advanced web crawling technology, employing specialized techniques tailored for the intricacies of the Dark Web. This comprehensive dataset comprises a significant volume of information, providing insights into various aspects of Dark Web activities. Structured for analytical purposes, it includes timestamped URLs, clean text, critical scores, parent URL of a web page (the URL of the page from which the current web page was linked or accessed) and other pertinent metadata. The dataset's organization and structure facilitate a systematic exploration, offering valuable information to comprehend and respond to cybersecurity threats within the Dark Web landscape. Using exploratory data analysis techniques, we aim at gaining insights from these data, thereby contributing to a comprehensive understanding of the nature of activity on the Dark Web. This examination of the contents of the Dark Web serves not only to demystify its hidden elements but also to shed light on the hidden aspects of the Internet and promote a more informed discourse on cybersecurity and digital security.

It is evident that the resulting dataset from this crawling comprise a graph. To uncover the information hidden in this dataset, we implemented a comprehensive workflow for sanitizing, storing, analyzing and visualizing the results of this analysis. Initially, unnecessary information (e.g. tags, and JavaScript) has been removed. Then the dataset was stored in plaintext in HDFS, Hadoop Distributed Filesystem and was processed using Apache Spark [14] enhanced by the GraphFrames module for graph processing. Then, we performed text analysis to identify the predominant keywords and we moved on to graph analysis, applying well known graph algorithms. We have made the strategic decision to utilize a big data engine for our workflow so as to be able to harvest large volumes of crawls. We specifically selected Apache Spark so as to utilize, in a consequent step, its stream processing capabilities.

In this paper we report the results based on a sample retrieved from the Dark Web. The results of our analysis indicate that the vast majority of the web sites found in the dataset are directly related with keywords directing to cryptocurrencies, as *"Bitcoin"* and terms associated with cyberthreats. The rest of this paper is organized as follows. Section 2, provides the building blocks of this paper. In Section 3 lie our analysis findings. In Section 4 we present works related to ours. Finally, we conclude in Section 5.

2 Preliminaries

In this section, Apache Spark is presented, which is the engine used for the analysis performed in this paper, together with a brief description of connected components and PageRank.

2.1 Apache Spark

Apache Spark [14] is an open source framework for memory-oriented big data processing. It is considered state-of-the-art, overtaking Hadoop MapReduce in many ways [9]: it is faster, easier to program, and also supports a variety of compute-intensive tasks, through rich APIs in a variety of languages (Scala, Java, Python, SQL and R). Its design is based on a data abstraction called Resilient Distributed Dataset (RDD). An RDD is a virtual distributed dataset consisting of partitions by defining *transformations* to their data. Transformations return new RDD objects representing the result of computations. But computations do not take place immediately, but only after specific commands called *actions* are defined. Thus, Spark is coined to follow *lazy evaluation*. RDDs may be transformed into *Dataframes* which represent data in named columns. In this work, we also utilized GraphFrames [5], an extension of Spark used for mixing graph and relational queries. GraphFrames operate on top of Spark's Dataframes allowing the seamless manipulation of graphs through Spark operators.

2.2 Connected Components

In graph theory, the connected component of an undirected graph is a subgraph in which any two vertices are connected by paths and which is not connected to any other vertices in the rest of the graph. For an undirected graph, the connected components are the largest possible connected subgraphs. However, in the case of a directed graph, the notion of connected components is more complex. In a directed graph, a connected component is said to be "strongly connected" if there is a direct path from any vertex to all other vertices in the component. A weakly connected component is a graph where a directed graph becomes connected if we replace all of its directed edges with undirected edges.

2.3 PageRank

PageRank [3] is a function that assigns a real number to each page in the Web. The concept is that the higher the PageRank of a page, the more "important" it is. While there is a basic algorithm for calculating PageRank, there are various modifications and extensions of the original algorithm that can result in different PageRank values for pages.

Keyword	Count
bitcoin	59737
hack	56399
btc	42394
encrypt	21836
script	19806
crack	13311
trojan	12287
malwar	10841
exploit	10073
ddos	7751

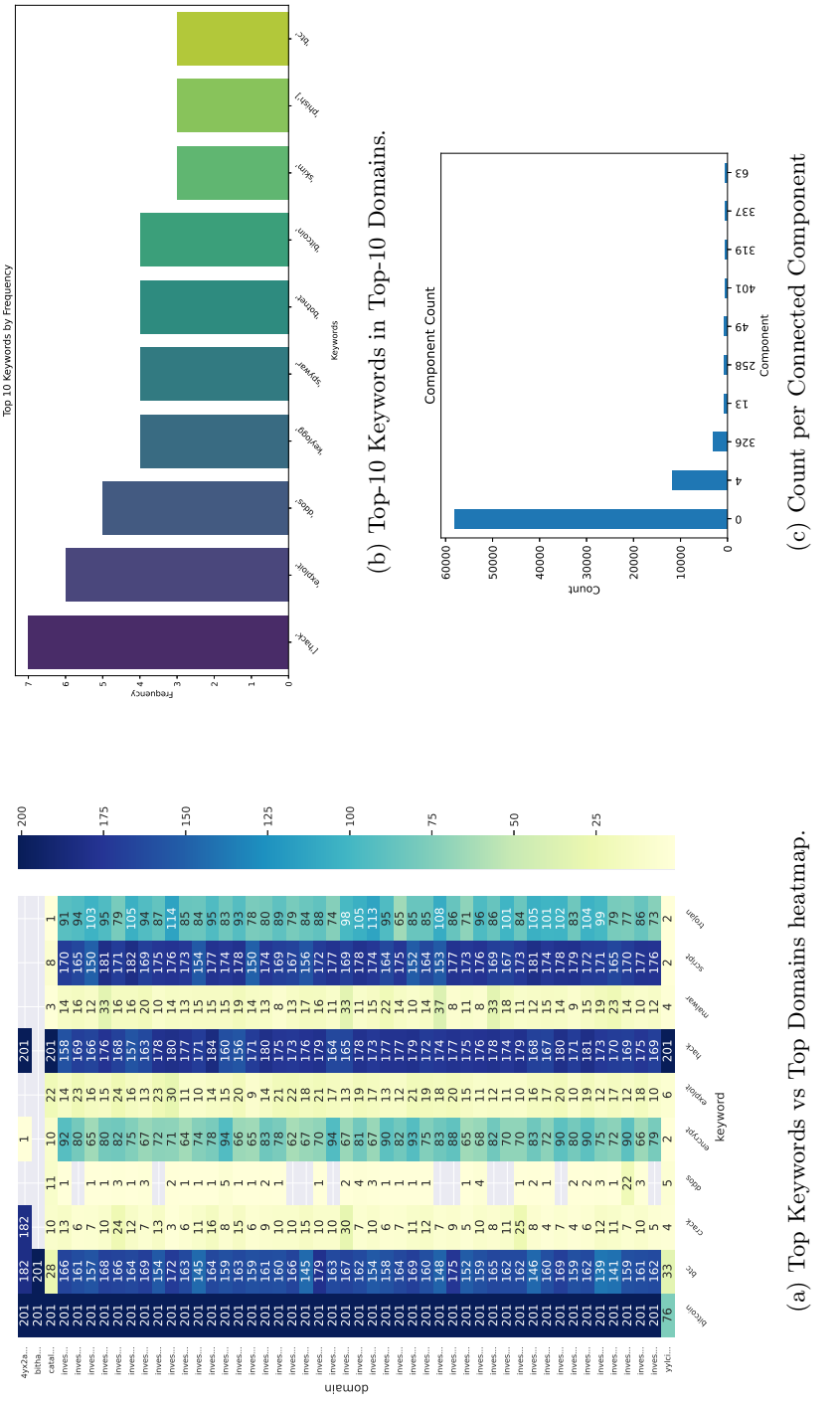
Table 1: Top 10 Keywords

Component Id	Top Keywords
0	bitcoin, btc
4	hack, encrypt, exploit, ransomware
13	bitcoin, breach, hack
49	bitcoin, hack
63	bitcoin, hack, vulner
258	bitcoin
319	bitcoin, btc
326	bitcoin, btc
337	bitcoin, btc
401	bitcoin, btc

Table 2: Top keywords / component.

URL	PageRank
http://tordexu73joywapk2txdr54jed4imqledpcvcuf75qsas2gwdgksvnyd.onion/directory	1921.4080088248134
http://t3g5mz7kgivhgzua64vxmu7ieyyoyzgd423itqjortjhh64lcvspyayd.onion	1806.6602692202455
http://vxmua4uvg7vp5ssnvx5gexrr2nxso3wwvjwagdub67vcombj4kf4i4qd.onion/catalogue.php?cat=hacking&page=1	1536.9789197954801
http://freewzvfroedixqklf6iqpwckoxsqah57qoqzfhrw6kaw2hjjbcf42id.onion/index.php?word=Bitcoin	1432.983030332785
http://vxmua4uvg7vp5ssnvx5gexrr2nxso3wwvjwagdub67vcombj4kf4i4qd.onion/catalogue.php?cat=hacking &page=2	1309.3594199101767
http://freewzvfroedixqklf6iqpwckoxsqah57qoqzfhrw6kaw2hjjbcf42id.onion	1218.361031038603
http://vxmua4uvg7vp5ssnvx5gexrr2nxso3wwvjwagdub67vcombj4kf4i4qd.onion/catalogue.php?cat=hacking	1117.1897539970873
http://vxmua4uvg7vp5ssnvx5gexrr2nxso3wwvjwagdub67vcombj4kf4i4qd.onion	996.6304339977689
http://5jvz3qy6rux3chaoft3sex22ef6l6uvkfqsavxd3pmej2jeotsckvxqd.onion	959.0582894990224
http://22tojepqmpah32fkeuurutki7o5bmb45uhmgzdg4l2tk34fkdafgt7id.onion	889.5988765667258

Table 3: Id's with the top 10 Pagerank values



(a) Top Keywords vs Top Domains heatmap.

(b) Top-10 Keywords in Top-10 Domains.

(c) Count per Connected Component

Fig. 1: Visualization of graph analysis results.

3 Analysis and Results

This section holds the analysis of the dataset used which consists of 85291 documents retrieved from the Dark Web. To access the Dark Web, we use a web crawler for this purpose, but we have first set up specific software, a TOR proxy [1], that grants us access to this network. This proxy acts as an intermediary between our crawler and the Dark Web. After establishing the connection to the Dark Web, we are all set to retrieve data. The high-level flow of the crawler is as follows: we start by providing a set of seed URLs to the crawler. These URLs are added to a queue, and then the system starts the crawling process by extracting the first URL from the queue and retrieving the content of the page. Several cleaning steps are performed to remove unnecessary information (HTML tags, JavaScript code, etc.). The next step is to extract the URLs contained in the retrieved page and add them to the queue. Lastly, we calculate a score based on the term frequency of the defined keywords.

For our analysis, we begin by identifying distinct domains in our data set. Then, we move on to investigating the existence of connected components within our graph, aiming at identifying distinct subgraphs within the web page network. After that, for each component, we applied PageRank to find the most influential web pages within each connected component. To gain a deeper understanding of each component, our analysis revealed the top domains and keywords, offering valuable insights into the distinctive content within each sector.

3.1 Text Analysis

Our first step is to perform text analysis. In the whole data set 4663 distinct domains were identified. Next, we extracted the Top-50 domains in terms of frequency of appearance in the data collection 90% (45 out of 50) of the URLs in these domains lead us to pages related to cryptocurrencies (contain the word invest within their URL) with information about their production as well as their purchase and sale. Out of the remaining five URLs, four directed us to the home directories of marketplaces like “Mike’s Grand Store” and “Darknet Home”, while the last one is no longer valid.

Table 1 presents the predominant keywords in our dataset, with a notable emphasis on Bitcoin in the first and third positions. This underscores the significance of cryptomining and trading in our dataset. The remaining entries in the table consist exclusively of words associated with cyberattacks. The Dark Web dataset is effectively visualized through the use of a heatmap, as illustrated in Figure 1a, which allows for the identification of prominent peaks associated with specific keywords. Noteworthy, among these are the terms “bitcoin”, “hack”, and “script”. The elevated frequencies of these terms suggest a substantial presence and discourse surrounding cryptocurrency, cyberthreats, and scripting languages within the Dark Web ecosystem. These elevated values may not only signify mere discussion but also imply the provision or seeking of services related to these keywords, thereby pointing towards potential illicit activities or specialized offerings.

3.2 Graph Analysis

Let us now move on and perform graph analysis on our dataset. We begin by executing PageRank, as implemented by Apache Spark and GraphFrames.

Overall PageRank Table 3 contains the top 10 pages of the dataset based on PageRank. The first, second, seventh, and ninth item of table 3 correspond to directories in four different Dark Web marketplaces: *Tordex*, *Deep Link Dump*, *Shops Dir*, *and Tasty Onions*. Then, entries 2, 4, and 6 correspond to categories of those directories, for example hacking etc. However, the links for the remaining entries indicate that they have either changed or are no longer operational. Based on these rankings, we can conclude that the directory is, as expected, more important than other links. Additionally, it is becoming apparent that interest is increasing in certain categories, namely Bitcoin and hacking.

Figure 1b illustrates a histogram for the Top-10 keywords in the Top-10 PageRank pages, visually capturing the dominance of key cybersecurity-related terms in the top-ranking pages, with "hack", "exploit", and "ddos" taking the lead. These terms underscore a strong focus on security vulnerabilities and cyberthreats. Following closely are terms like "spyware", "botnet", "bitcoin", and "phish", offering insights into a diverse range of topics, including digital security and cryptocurrency.

Connected Component Analysis Figure 1c illustrates the sizes of the Top-10 connected components across the dataset. The horizontal axis stands for the connected component id, as identified by executing Spark's connected components algorithm. The vertical axis stands for the count of the number of members of the respective subgroup. As we can see *Component_ 0* is the largest one, featuring a substantial count of 5800 members, illustrating its prominent role within the dataset. Then, *Component_ 4* follows with 1200 members, exhibiting a noteworthy but comparatively smaller presence. *Component_ 326*, depicted with less than 1000 items, signifies a specific and more confined subset of the data. The rest of these subgroups exhibit significant smaller sizes.

Now, we continue our analysis with a very interesting experiment to further investigate the characteristics of each such component. We have applied PageRank within each of these Top-10 connected components to identify the Top-10 most important nodes, according to PageRank results. Then, we identified the 10 most important keywords for each case, generating a set of heatmaps for easier visual inspection[3]. The results are summarized in Table 2. Next, we present the detailed results of this evaluation starting with Component_0.

Component_0. This is the largest component of all with 6000 members. After applying PageRank, we performed a keyword count In the heatmap, we notice the high usage of the words "bitcoin" and "btc" by all the top domains

[3] Detailed results and the code used to produce them may be found at: `https://github.com/klebash/Dark_Web_Exploratory_Analysis`

of the component. Then, all the other top keywords of the component are related to cybersecurity, with the main purpose of providing services in finding vulnerabilities in the systems to be attacked("hack", "crack", "encrypt", "exploit", "script", and "vulner", referring to "vulnerability") as well as the creation and transmission of viruses ("malwar", referring to "malware") to extract sensitive information.

Component_4. Here, The top-ranked domain in PageRank importance is strongly associated with keywords "encrypt" and "exploit". Top-4 PageRank ranked domains are strongly related to the keyword "ransomware". Top-9 PageRank ranked domains are related to the keyword "hack" It is evident that all these keywords pertain to cybersecurity, vulnerabilities, and attack methods.

Component_13. In Component_13, there is no particular keyword spanning along Top-10 PageRank-ranked sites. Nevertheless, all Top-10 keywords, feature similar frequencies in the top ranked domain, which contains the "Hidden Wiki", a forum where you can upload anonymous content and contains articles that explain in detail cybersecurity principles. As such, keywords as "bitcoin", "breach" and "hack" appear with high frequency.

In **Component_49**, the top-ranked page refers to the homepage of "Hacktown", a site where users can acquire skills from an ex-cybercriminal in hacking and fraud. Here, all Top-10 keywords appear, while there is sparse appearance in other sites. The most common keyword in all sites here is "bitcoin", while Top-4 sites also feature the keyword "hack" as common.

For **Component_63** There is no uniformity in keywords among the top domains. Top-ranked URLs are searches on *Tordex* which is a search engine for Tor. Here, the most common keywords are "hack" and "vulner", while in all URLs, there are appearances of "bitcoin", even in small number of occurrences.

In **Component_258** top-ranked pages are purchases of illegal services. Most URLs have occurrences of "bitcoin". The situation is similar for **Component_319**. However, here, all URLs have references to "bitcoin" and Top-9 to "btc", while the 10th URL also exhibits 16 occurrences of "hack".

Component_326, on the other hand, exhibits homogeneity. All URLs have high frequencies of "bitcoin", and all but Top-1 of "btc". The situation is similar for **Component_337** and **Component_401**. This is primarily due to the fact that the websites within these components are engaged in cryptocurrency mining and trading activities, rather than utilizing Bitcoin as a transactional currency. This conclusion is further supported by the most visited URLs, as determined by their PageRank scores.

4 Related Work

Data collection and analysis of the content of the Dark Web is a topic that has concerned the scientific community before. In [4] there is a comprehensive study on Dark Web mining. In [7], authors propose a framework for conducting more effective investigations on Dark Web marketplaces. The framework includes various techniques and approaches to gather information and intelligence on

marketplace vendors operating on the Dark Web. Recently, Georgoulias et al. [6] not only confirmed the abundance of cybercrime products in the Dark Web, but they also identified their availability in low prices. These marketplaces often feature security mechanisms for their users, making data harvesting difficult. Wang et al. recently presented a review of such mechanisms [13]. A review of the user experience in the Dark Web may be found in [11], while threat types are described in [12]. Methodologies and techniques for analyzing and visualizing the content and network properties of the Dark Web are presented in [10]. Tools for data gathering and analysis may be found in [2].

5 Conclusions and Future Work

In this paper, we performed an analysis of the structure and the content of a large number of interconnected Dark Web sites. We identified the proliferation of top keywords primarily focused on "Bitcoin" and terms related to cyberthreats illustrating a dynamic environment in which cryptocurrencies and cyberthreats coexist. Our next steps include analysis of even larger Dark Web crawls.

Acknowledgments. This work has received funding from the European Union's Horizon Europe research and innovation programme under the Ceasefire project with Grant Agreement No 101073876.

References

1. Tor project. `https://www.torproject.org/`
2. Ball, M., Broadhurst, R., Niven, A., Trivedi, H.: Data capture and analysis of darknet markets. SSRN Electronic Journal (2019)
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems **30**(1-7), 107–117 (1998)
4. Chen, H.: Dark web exploring and data mining the dark side of the web. Springer (2012)
5. Dave, A., Jindal, A., Li, L.E., Xin, R., Gonzalez, J., Zaharia, M.: Graphframes: an integrated api for mixing graph and relational queries. In: Proceedings of the fourth international workshop on graph data management experiences and systems. pp. 1–8 (2016)
6. Georgoulias, D., Yaben, R., Vasilomanolakis, E.: Cheaper than you thought? a dive into the darkweb market of cyber-crime products. In: Proceedings of the 18th International Conference on Availability, Reliability and Security. pp. 1–10 (2023)
7. Hayes, D., Cappa, F., Cardon, J.: A framework for more effective dark web marketplace investigations. Information **9**(8), 186 (2018)
8. Pantelis, G., Petrou, P., Karagiorgou, S., Alexandrou, D.: On strengthening smes and mes threat intelligence and awareness by identifying data breaches, stolen credentials and illegal activities on the dark web. Proceedings of the 16th International Conference on Availability, Reliability and Security (2021)
9. Shanahan, J.G., Dai, L.: Large scale distributed data science using apache spark. In: The 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2323–2324. ACM (2015)

10. Takaaki, S., Atsuo, I.: Dark web content analysis and visualization. Proceedings of the ACM International Workshop on Security and Privacy Analytics (2019)
11. Tazi, F., Shrestha, S., De La Cruz, J., Das, S.: Sok: An evaluation of the secure end user experience on the dark net through systematic literature review. Journal of Cybersecurity and Privacy **2**(2), 329–357 (2022)
12. Türen, S.H., Islam, R., Eustace, K.: Analysing the threat landscape inside the dark web. In: Emerging Trends in Cybersecurity Applications, pp. 95–122. Springer (2022)
13. Wang, Y., Arief, B., Hernandez-Castro, J.C.: Analysis of security mechanisms of dark web markets (2024)
14. Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I.: Spark: Cluster computing with working sets. In: 2nd USENIX Workshop on Hot Topics in Cloud Computing, HotCloud'10. USENIX Association (2010)