IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 9, 2021, accepted January 26, 2021, date of publication January 29, 2021, date of current version February 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3055532

Exploring the Topological Properties of the Tor Dark Web

ABDULLAH ALHARBI[1], MOHD FAIZAN[2], WAEL ALOSAIMI[1], HASHEM ALYAMI[3], ALKA AGRAWAL[2], RAJEEV KUMAR[2,4], AND RAEES AHMAD KHAN[2], (Member, IEEE)

[1]Department of Information Technology, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia
[2]Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Lucknow 226025, India
[3]Department of Computer Science, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia
[4]Department of Computer Application, Shri Ramswaroop Memorial University, Barabanki 225003, India

Corresponding authors: Abdullah Alharbi (amharbi@tu.edu.sa) and Raees Ahmad Khan (khanraees@yahoo.com)

This research was supported by Taif University Researchers Supporting Project number (TURSP-2020/231), Taif University, Taif, Saudi Arabia.

ABSTRACT The web graph can be used to get insight into the internal structure and connectivity of the Tor dark web. This paper analyzes the internal structure of the Tor dark web graph and examines the presence of bow-tie structure as found in the World Wide Web. The web graph is generated from the data collected by the Python crawler customized to scrape data from the Tor dark web. Each of the nodes in the graph represents an individual Tor hidden service, and an edge denotes the hyperlink from one hidden service to the other. Various graph metrics are then computed and analyzed for both directed and undirected graphs using the Python NetworkX package. It was found that most of the nodes of the graph have in-degree and out-degree less than ten. The presence of power-law in degree distribution could neither be confirmed nor denied. The Tor web graph is sparse with a few connected pairs of nodes. Like the surface web, the dark web can also be decomposed into a bow-tie structure though with small component sizes. Several important and well-known websites on the surface web have incoming links from the dark web. Moreover, the Tor network also shows the characteristics of small-world and scale-free networks.

INDEX TERMS Bow tie, dark web, graph analysis, Tor, web graph.

I. INTRODUCTION

A web graph can be described as a collection of vertices and edges where a single vertex depicts a web page, and an edge between two vertices is a directed hyperlink from a web page to the other. All the vertices and edges belong to the World Wide Web or the Internet. The study of web graph may help in identifying the underlying structure of the Internet of how the different web pages are linked to each other. They help in developing efficient data mining techniques and better crawling strategies. The exponential growth in the size of the World Wide Web has attracted the scientific community to investigate its graph structural properties with various aspects [1]–[4].

Dark Web is a portion of the Internet that can only be accessed using sophisticated routing techniques. The anonymity provided to the dark web users have been misused to carry out illicit activities online [5]. The law enforcement agencies have expressed concerns regarding the criminal usage of the dark web [6]. Moreover, it also becomes challenging given the ever-changing structure and high churn in the dark web ecosystem [7]. There is a need to study the structure of the dark web to get insight into the criminal activity on the dark web. To the best of our knowledge, only limited work has been done on the graph structure of the dark web as the majority of studies have focused on the nature of content present in the dark web and its legal acceptability.

This paper has tried to fill this gap by exploring the directed and undirected Tor web graph at the domain level. The data has been collected from the dark web using the tailored made web crawler. The Python NetworkX package has been used to generate the graphs from the collected data. The web graph consists of 48,174 vertices or nodes and 103526 edges. Various graph metrics have been calculated and analyzed for the graph to identify the importance of different hidden services. The connectivity within the graph and its structural shape are also investigated. Also, the connectivity to the surface web from the dark web is studied.

The rest of the paper is organized as follows: Section II briefly describes the dark web and the Tor. Section III contains the related work on the surface web and the dark web. Section IV describes the methodology employed for the

The associate editor coordinating the review of this manuscript and approving it for publication was Zhangbing Zhou.

collection and analysis of data. Section V reports the result obtained for various metrics. A discussion about the results is done in Section VI. Finally, Section VII concludes the study with future prospects.

II. ABOUT DARK WEB AND TOR

The Internet web pages that regular user access from the web browser is called the surface web. The part of the Internet that exists below the surface web is commonly referred to as the deep web. The portion that lies even below the deep web is called the dark web. It is an encrypted network and uses the Onion routing [8] scheme to connect to web services. The Tor browser[1] is used to access the dark web. The user of the Tor browser remains anonymous while surfing the dark web. It is also known as the Tor network. The websites in the Tor dark web are often referred to as the hidden services. The Uniform Resource Locator (URL) of a hidden service is a 16 character length string of random alphabets and digits, which ends with a top-level domain called *.onion*.

The dark web is being used for both good and corrupt activities. Users with good intentions can leverage this anonymous platform to express their thoughts and views openly without any censorship [9]. The law enforcement organizations and government bodies may use the dark web to secretly carry out their missions [10]. Moreover, such platforms are popular among the people of oppressive regimes where Internet access is highly controlled [11].

The downside of the dark web platform is that it is being misused by criminals and fraudsters. Many of the illicit trades that were previously carried out on the streets have now been shifted to the dark web [12]. The sale of illegal drugs being one of the other grey activities carried out on the dark web [13]. Other controversial activities include child abuse, hate, violent and gory content, credit card frauds, pirated software, unethical hacking guides, etc. [5], [14]. These services and content may be available individually on single hidden services or they may also be available under a single roof commonly referred to as cryptomarkets. Silk Road, Dreammarket, Alphabay were the popular cryptomarkets on the Tor dark web [15].

The two opposite usage of the dark web platform creates a state of dilemma for law enforcement agencies to undertake a strict action [6]. Hence there is a need to get insight into the internal structure of the dark web, which may help in the better monitoring of activities being carried out there. The web graph of the dark web may help in identifying the central nodes that hold the network structure. It may also help uncover how the network structure supports the illicit activities.

III. RELATED WORK

This section begins with the description of related work on the analysis of the Tor dark web graph followed by the discussion of some significant contributions related to the analysis of the web graph of the World Wide Web/Surface web.

The research work aims at exploring the topological structure of the dark web is limited. A critical study is by Bernaschi *et al.*, 2017 [16], where they collected data using the BUbiNG crawler for the analysis of graphs at the page level, host level, and service level. They have reported graph metrics like degree distributions along with the identification of connected components of the graph and their importance in the integration of the entire graph structure. Both the directed and undirected graphs were analyzed. They also performed the semantic analysis of their data to relate it with their findings of the topology of the web graph.

There are similar works that are carried for the surface web in the last two decades. A few of the early studies to determine the topological properties of the web graph were by Kumar *et al.*, 1999 [17] and Barabási *et al.*, 2000 [18] where the former reports the inverse polynomial distribution being followed by the in-degree and out-degree distributions while the latter claimed the power law. Kleinberg *et al.*, 1999 [1] proposed in their paper two algorithms for web graph, which they ran on their data set to obtain measures for better search. Based on their observations, they also propose a family of random graph models that suited their measurements. Broder *et al.*, 2000 [2] analyzed the web graph generated from a large dataset of 200 million pages with 1.5 billion links provided by AltaVista crawl. They found an important result that the web graph bears a structure similar to that of a bow tie having six components. According to them, the in-degree and out-degree distributions follow the power-law and affirm it as the underlying property of the Web. They also claim that the measurements of the graph properties do not have any effect due to the change in crawled data.

A similar study by Donato *et al.*, 2007 [19] conducted their experiment on the data set provided by the WebBase project. They reconfirm the presence of power-law exhibited by in-degree distribution as observed in previous studies. They also found that there exists a small correlation between the page rank and in-degree distribution. The bipartite cliques were also found in their analysis of the web graph. They made a comparison between their WebBase crawl and the Altavista crawl use by Broder *et al.*, 2000 [2] concerning the size of the components of bow-tie structure and found differences in both datasets. This may be due to the time at which the two crawls were performed (1999 and 2001).

Lehmberg *et al.*, 2014 [3] analyzed the surface web graph aggregated at the page level domain (PLD). Their coverage of more than three billion pages and their hyperlinks was collected in 2012. Once again, the in-degree distribution was following a power law with exponent 2.4 in their findings. They confirm the presence of a bow-tie structure in the PLD graph as set forth by Broder *et al.*, 2000 [2], for the page graph. They also categorized the websites into different topics and discovered the connectivity between the website belonging to different categories. Based on the overall observations, they suggest a hypothetical structure for a PLD graph called

[1] https://www.torproject.org/

the Two-Layer Model which consists of a Low Degree Layer that contains sparsely connected websites and a High Degree Layer containing densely connected websites with high in-degree.

Meusel *et al.*, 2015 [4] performed graph analysis at different aggregation levels on the data set collected by the Common Crawl Foundation in 2012. Comparing their results with the previous studies, they reported a significant increase in the average in-degree and the connectivity of the page graph. As the existence of a large strongly connected component in page and PLD graph was confirmed by the previous studies discussed above, its existence in the remaining host graph was also confirmed by them in their work. The distance-based metric for the graphs was also computed at three different aggregations levels.

In a slightly different study, Serrano *et al.*, 2007 [20] performed a comparative study of different crawl mechanisms and their effects on the various graph-theoretic properties. They put forward the need for a framework for the analysis of sampling biases that occurs in different crawl techniques. According to them, it will help in determining the exact structure of the Web, which varies due to the change in the crawling mechanism.

The extraction of data from the web is usually referred to as Web Information Extraction. The method of collection of data from the websites is an important task for a good analysis of the results. The process of extraction of data from the dark web is similar to that on the surface web as they both are built using a hypertext markup language (HTML). Several existing works have surveyed the tools and techniques for extracting data from the surface web and deep web [36]–[38]. The data extraction from the web can be performed either manually or using fully automatic methods. Automatic extraction can be performed to scrape a chunk of content from websites like news feeds or blogs. In other scenarios, the websites can be automatically extracted to retrieve frequent patterns or sequences to label the websites into specific categories [36].

A survey study has identified two distinct approaches for extracting data from the web [37]. The first approach relies on the structural properties of the web pages that are built upon HTML. Such techniques consider the tree-like structure of the HTML tags and extract required data embedded within the tags. These methods give better performance on the web pages having similar HTML markup but may not be effective on web pages with different HTML structures. The second approach is based on machine learning techniques where a model is trained to predict the label of a particular web page. In practice, machine learning methods usually need a large amount of training data to produce a good performance. However, such a method may not be able to predict the new instances that were not present in the training set.

To overcome the challenges associated with web data extraction, an extractor is presented based on supervised and unsupervised methods to extract the specific product description from the websites [39]. An unsupervised technique called visual validation was combined with a supervised classifier to produce a versatile extractor that works on a variety of websites. The extractor was implemented on a publicly available corpus. The results on the corpus showed that the visual validation technique can improve the extractor performance when trained on the visual features.

IV. METHODOLOGY

This section defines the properties of the graph that shall be explored in the Tor network and the procedure for collecting the data used for the construction of the Tor web graph.

A. REVISIT TO THE GRAPH TERMINOLOGY

Some basic graph terminology used in this paper is briefly described in this section. Readers familiar with the graph terminology may skip this part.

1) GRAPH

A graph or an undirected graph G denoted as G = (V, E) is a collection of two sets V and E where V is a finite nonempty set of vertices, also called nodes, and E is a set of edges or arcs or lines that connects the vertices. A directed graph or a digraph is one in which there is a specified direction between vertices as represented by edges. An edge (A, B) represents a link starting from vertex A and ending at vertex B.

2) PATH

A path $v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow \cdots\cdots v_n$ is a sequence of edges between two vertices v_1 and v_n such that $v_i \rightarrow v_j$ represents an edge from the vertex v_i to v_j.

3) DEGREE

A vertex degree is the total number of edges that are incident on it. For a vertex in the directed graph, the in-degree is the total number of incoming edges to the vertex, and the out-degree is the total number of outgoing edges from the vertex. The degree of a vertex in a directed graph is the sum of its in-degree and out-degree.

4) CONNECTED COMPONENTS

A connected component S in an undirected graph is a set of vertices such that each vertex in it is reachable from every other vertex in S.

5) STRONGLY CONNECTED COMPONENTS

A strongly connected component (SCC) of a directed graph is a set of vertices S such that there is a directed path between any two vertices (x, y) of S. A weakly connected component (WCC) of a directed graph is a set of vertices S such that there is a path between any two vertices (x, y) of S ignoring the direction.

6) CENTRALITY

A centrality of the vertex is a quantitative measure of the importance of that vertex in the graph. A range of centrality measures exists.

Note that both nodes and vertices are used interchangeably in this paper. For convenience and better readability, the Tor Hidden Services web graph will be referred to as a web graph or graph for simplicity. Also, the Tor Hidden Services are synonymously used with websites.

B. DATA COLLECTION

This section reports the crawling process applied to gather the data, followed by its pre-processing. Then the description of the extraction process of links for graph generation is given.

The data for the study was collected by the custom made web crawler. The crawler was coded in Python, which makes connections to the Tor Hidden Services using SOCKS proxy [21]. The crawler was supplied with the initial onion domains or seeds from the publicly available directory of Tor links [22] from the surface web. It scrapes each of the seeds and stores the new links found in a separate file. The fresh links were again crawled to discover further new links. The process of scraping new links is successively repeated two more times until no new links can be found. The crawler did not scrape the hidden services that require user login or were behind subscription pay-walls. At the end of the crawling session, 48,174 online hidden services were left for further processing.

As per the statistics of the Tor Project Inc., there are more than 100,000 hidden services available online on any given day. However, our crawler could only capture around 48,000 hidden services online. The gap in the numbers of hidden services may be attributed to the chatting platforms on Tor like TorChat [23], where each of the participating users is assigned a unique 16 character *.onion* address.

This paper investigates the structure of the dark web at the domain level, so all the sub-domains of a particular domain are aggregated under it and are denoted as a single individual node of a graph. To be specific, consider a hidden service domain *XYZ.onion* having two sub-domains *en.XYZ.onion* and *ch.XYZ.onion*, both of these sub-domains and their internal web pages are treated as a single node of graph identified by their domain name. Hence any edge from node X to node Y in the graph represents that there is a hyperlink within the web page of domain X pointing to a web page contained in domain Y.

For the construction of the graph, the collected domain web pages are searched for the links in HTML <*a*> tag to other pages and saved in a separate file with the help of Python code. The hyperlinks that point within the web pages of the same domain are discarded. Once the links of all the domains are obtained, an adjacency list for the graph has been generated. This adjacency list is supplied to the Python NetworkX library for the generation of the directed web graph. The undirected version of the graph was also created from the directed graph.

Upon completing the process described above, the graph was obtained having 48,174 nodes and 103526 edges. Each node in the graph represents a unique onion domain, and an edge represents a hyperlink between two different domains.

V. RESULTS AND ANALYSIS

This section reports the results obtained from the collected data for the different graph properties and their analysis.

A. DEGREE DISTRIBUTIONS AND PAGERANK

The in-degree distribution of nodes (having an in-degree value of less than 50) is presented in Fig. 1. In Fig. 1, it can be seen that ∼7% of nodes are source i.e., they have zero in-degree. Followed by this are ∼39% of nodes having in-degree 1. On adding the nodes having in-degree up to 10, we get ∼97% of total nodes. This result is in line with the previous study [16] about the hidden nature of Tor web pages that are difficult to discover as they have few incoming links. In our analyses, we found that only seven nodes have in-degree higher than 100, of which the maximum in-degree being 184. Fig. 2 shows the in-degree distribution on the log-log scatter plot. There are some spikes present on the plot as the in-degree increases.

The existing work on the surface web has reported the presence of the power law in the degree distribution [24].In Fig. 2, the distribution seems to be following power law as the points tend to fall on a straight line at the beginning. To statistically confirm the presence of the power law, we have applied the methodology suggested by Clauset *et al.* 2009 [25] which utilizes the Kolmogorov-Smirnov goodness of fit test. The p-value obtained was greater than 0.1 which indicates the presence of the power law.

The out-degree distribution of nodes (having an out-degree value of less than 50) is shown in Fig. 3. Around 75% of nodes are sink having zero out-degree and 98% of nodes have out-degree less than or equal to 10. We found that 31 nodes have out-degree greater than 100 and out of them, nine nodes have out-degree greater than 1000. The maximum is 2846 links, thus covering a large portion of the web graph. All of the highest out-degree hubs are the websites offering directory services and Wikis. Although these nodes have a large number of outgoing links, they themselves have less than ten incoming links making them hard to find. The most noticeable feature that came across was the presence of isolated nodes which renders the web graph disconnected. A total of 3006 nodes were disconnected from the entire web graph having no incoming or outgoing links. Fig. 4 shows the out-degree distribution on a log-log scale. The distribution does not seem to be following the power law as the points are scattered haphazardly.

More than 95% of nodes have their in-degree and out-degree lies between 1 and 10, which makes the web graph a sparse one. Most of the edges are clustered around a few high out-degree hubs.

The largest in-degree and out-degree node were of comparable size in the surface web [4], but the same could not be held in the dark web as evident in the above discussion. The size of the largest out-degree node in the dark web is nearly fifteen times more than the largest in-degree node.

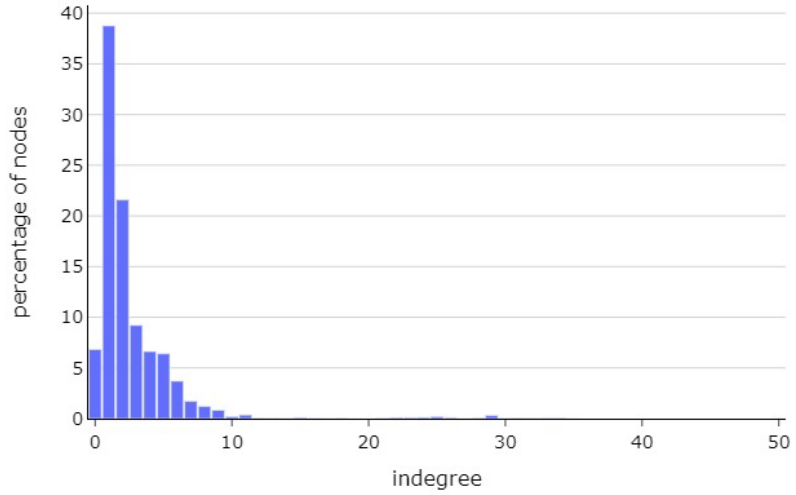

FIGURE 1. In-degree distribution.

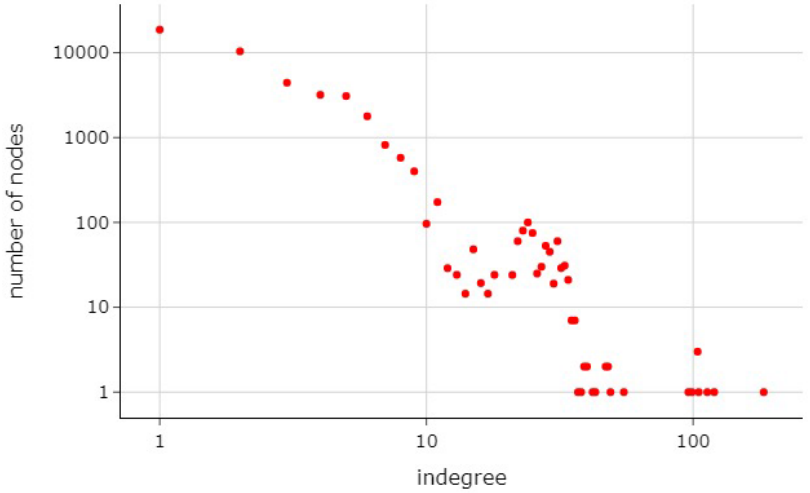

FIGURE 2. In-degree distribution (log-log scale).

Consequently, the in-degree distribution decays slowly than the out-degree distribution.

Fig. 5 depicts the PageRank distribution of the nodes. PageRank is used to measure the significance of a page in a graph [26]. It gives the probability value between 0 and 1 for each of the nodes in a directed graph. The damping factor taken for calculating PageRank is 0.85. The node with a higher PageRank value will have a higher chance of being accessed from any random node in a graph. The correlation coefficient between the PageRank and in-degree of a node is 0.88 which indicates a strong positive linear relationship between them. The top four ranked nodes are the same in terms of both the PageRank and the in-degree value. This finding is also comparable to the PLD graph of the surface web, where the top-ranked PLD is common in both PageRank and in-degree values [4].

As evident from Table 1, node #1 has higher PageRank than node #2 even though it has fewer numbers of incoming links. Since node #1 offers Bitcoin-related service so it may have incoming links from other relevant websites on the dark web having good PageRank, which subsequently contributes to its higher value. Besides, the high value of node #2 indicates the popularity of the Dream Market Forum it hosts. It should be noted that Dream Market is one of the most famous marketplaces on the Dark Web [27]. The out-degree does not contribute to the PageRank value.

B. EIGENVECTOR CENTRALITY

Centrality metrics are used in graph theory to identify the prominent vertices in a graph. In this paper, eigenvector centrality is used to locate important nodes in the graph. It relatively assigns a score to each of the nodes present in

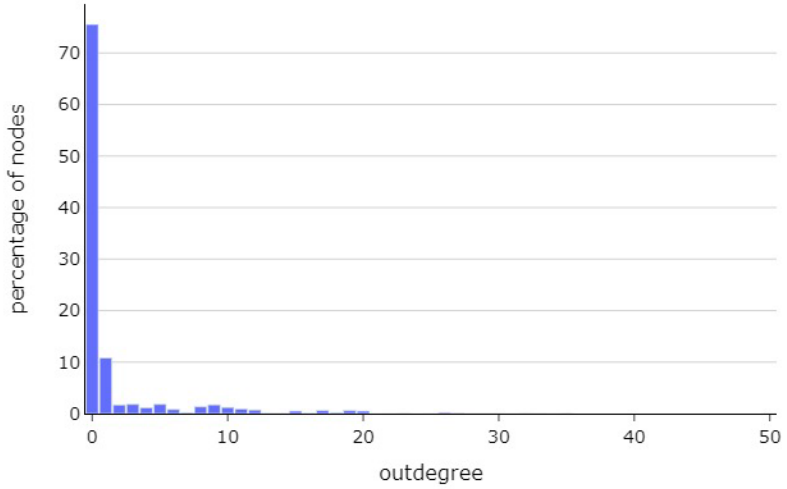

FIGURE 3. Out-degree distribution.

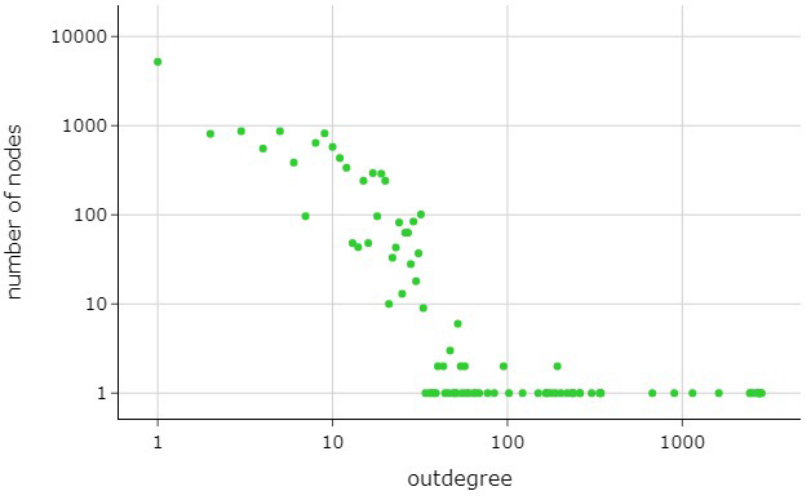

FIGURE 4. Out-degree distribution (log-log scale).

TABLE 1. Top 4 PageRank nodes.

S.No.	PageRank	In-degree	Out-degree	Description
1.	0.037	120	0	100x Your Coins in 24 Hours - Officially Hidden Service Anonymous
2.	0.024	184	0	Dream Market Forum
3.	0.012	113	1	Dream Market Login - Featured anonymous marketplace
4.	0.009	105	7	Dream Market Login - Featured anonymous marketplace (Mirror)

the graph based on its connection to the other nodes in the same graph. A node is accredited with a high score having connections to other high scoring nodes than the node having an equal number of connections but to the low scoring nodes. More specifically, a node will have importance if its neighboring nodes are important. Fig. 6 shows the distribution of eigenvector centrality of the nodes of the undirected graph.

From Fig. 6, it can be seen that nearly 70% of nodes have their eigenvector centrality value lies in the range 1E-02 to 2E-02. The percentage of nodes rapidly falls as the value of eigenvector centrality increases. Table 2 shows the five

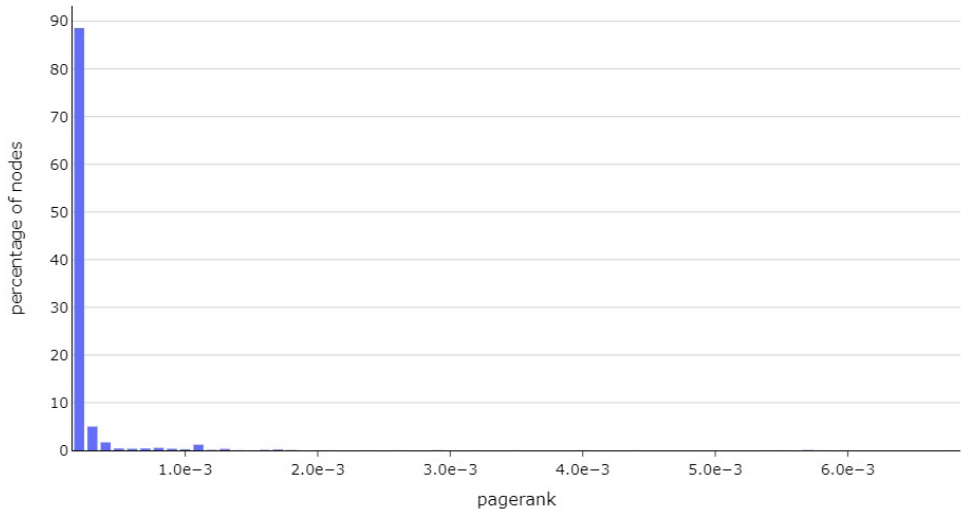

FIGURE 5. PageRank distribution.

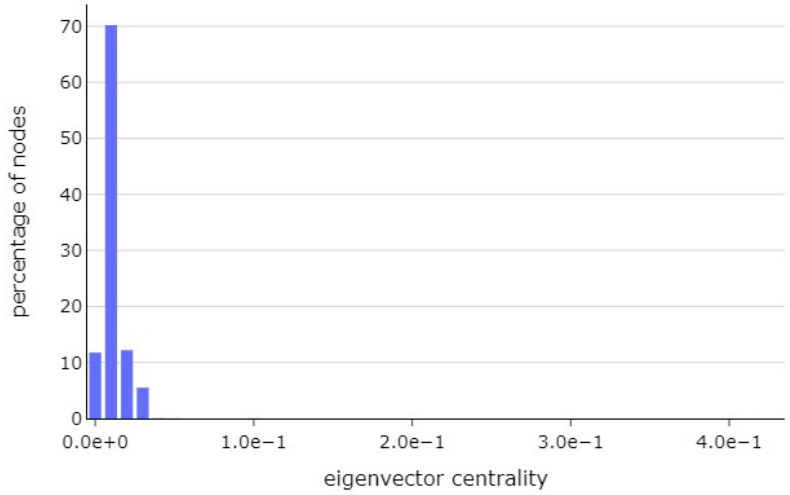

FIGURE 6. Eigenvector centrality distribution.

TABLE 2. Top 5 eigenvector centrality nodes.

S.No.	EGV Centrality	Out-degree	In-degree	Description
1.	4.713E-01	2846	0	CB3ROB Tactical Data Services - TOR Darknet site listing
2.	4.50E-01	2731	2	The Hidden Directory
3.	4.182E-01	2474	9	The onion crate - Tor hidden service index
4.	3.787E-01	2150	4	Jack's Tor Hidden Links
5.	3.524E-01	1684	4	Dark Web Hub

nodes with the highest eigenvector centrality and their degree metrics.

Unlike PageRank, which was not affected by the out-degree, the eigenvector centrality entirely depends on the out-degree and is unaffected by the in-degree. As discussed above, 98% percent of nodes have out-degree less than 10 which consequently leads to smaller eigenvector centrality for the majority of the nodes. Moreover, only a handful of nodes with a large number of out-going links have bigger eigenvector centrality values (as shown in Table 2). The correlation coefficient between eigenvector centrality and out-degree is 0.92 which confirms

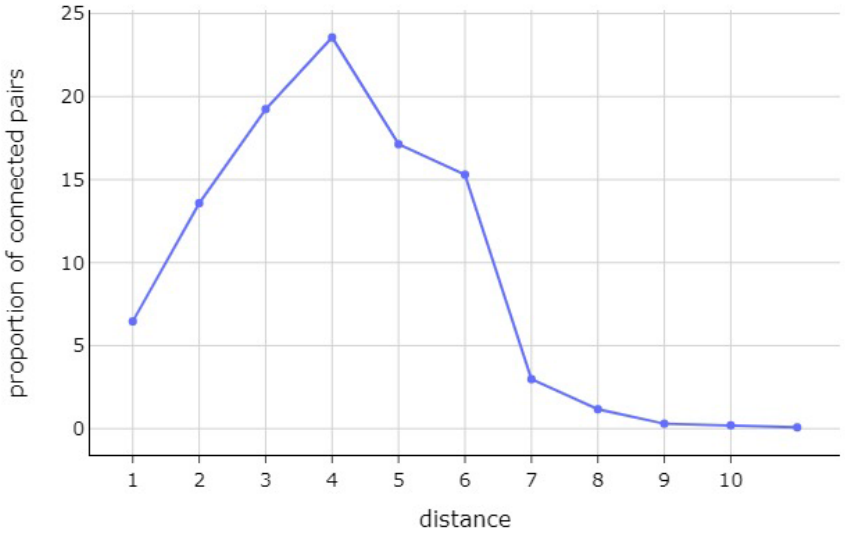

FIGURE 7. Distance distribution of connected pairs.

that the two variables tend to have a perfect linear relationship.

C. DISTANCES

Of all the possible pairs of nodes in the directed graph, only 2.3% of pairs were connected having a directed path between them. Such a low number of connected pairs in the graph once again affirm its sparse nature. The shortest path length ranges between 1 and 12 with an average distance of 4.32, which means one can travel across the connected pairs by passing through at most three nodes. Despite low internal connectivity, the average distance is nearly equal to the average distance of the surface web which is 4.27. However, unlike the dark web graph, the number of connected pairs on the surface web was above forty percent [4]. Fig. 7 shows the distribution of the distance between connected pairs with the percentage of connected pairs. The connected pairs can be seen clustered around the average value.

D. CONNECTIVITY

The presence of isolated nodes makes the web graph disconnected. To examine the connectivity, all the isolated nodes have been removed from the graph and rerun the algorithm to identify the connected components in the directed graph as well as in the undirected version obtained from the directed graph. The undirected graph was found to be disconnected even after the removal of isolated nodes. It contains eight connected components. The largest component consists of 45089 nodes followed by 57 nodes in the second largest component. The sizes of the remaining components were below ten nodes. If we exclude the isolated nodes, the largest component covers nearly 100% of nodes; thus, there is a strong probability of traversing a complete graph if any of the nodes of the component can be reached. Similarly, in the surface web graph, the largest component covers more than ninety percent of the websites [4].

However, upon removing the top five highest out-degree nodes (these five nodes contain the bridges of the graph whose removal increases the number of connected components.) from the graph, the largest component size get reduced to nearly half of the nodes covering only about 48% portion of the graph. This leads to an important result that although removal of the top five nodes does reduce the component to its half, the internal connectivity remains intact irrespective of the fact that all the remaining nodes left in the component were of out-degree less than 10. It shows the robustness of the graph structure that does not entirely disintegrate even after removing all the largest hubs. On the other hand, there was no significant reduction in the size of the largest component upon the removal of the vertex having the largest in-degree (i.e., 184).

Coming to the directed graph, four strongly connected components S1, S2, S3 and S4, were present of a considerable size of 1796, 83 and 27, 16 respectively, and the remaining SCC were less than 10 in size. In this case, it differs from the surface web, where the largest strongly connected component spans across more than fifty percent of the websites [4]. Unlike in the undirected graph, the removal of the highest in-degree and out-degree nodes does not have any effect on the size of S3 while S1, S2 and S4 get reduced to 59, 21 and 5 nodes respectively. In the case of S3, this may be attributed to the fact that all the 27 nodes of it are the clone of the same marketplace running under the different domain names. This reveals the introvert nature of Tor Hidden Services, where the related services are internally connected to each other, with only a few outgoing links to other services. The reason for the connectivity of S1 is the presence of many of the Directory/List/Wiki services in it which will obviously have links to the majority of other services.

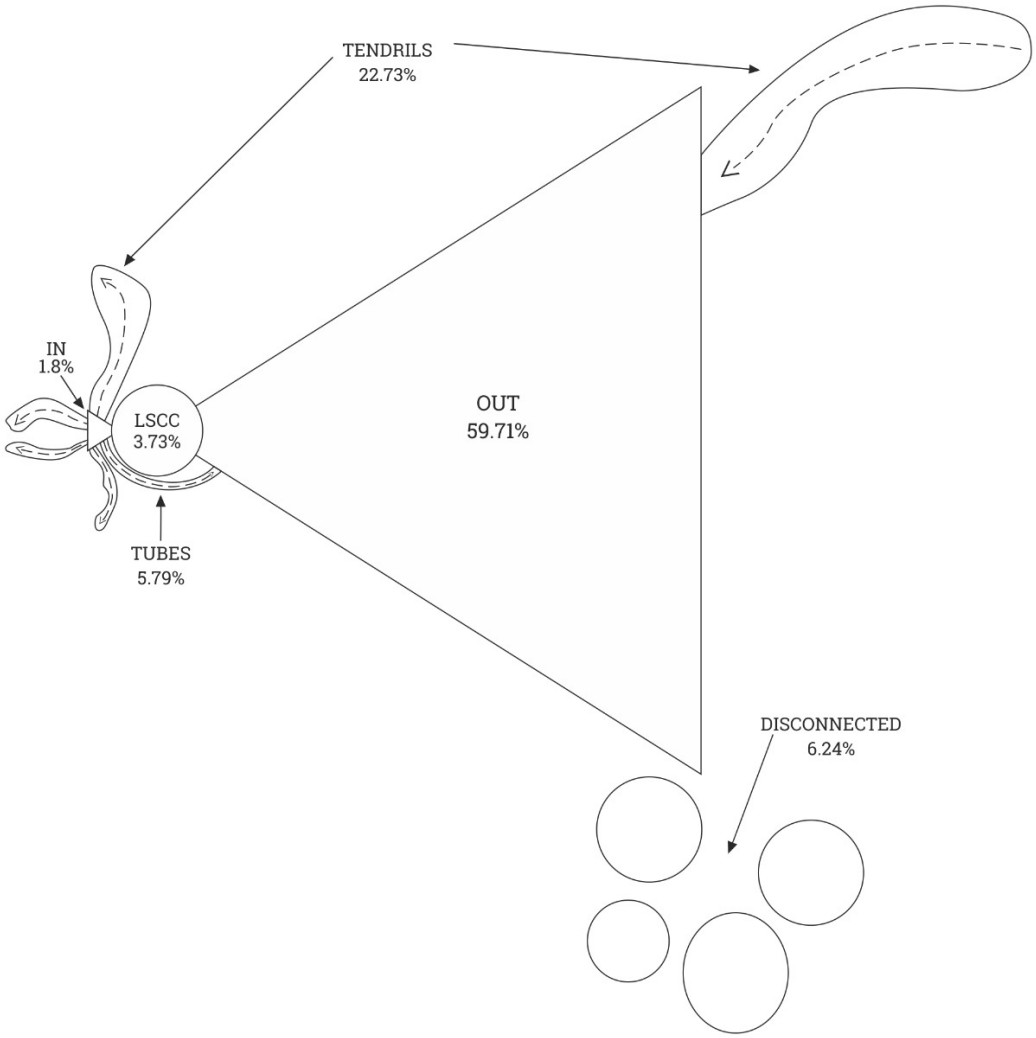

FIGURE 8. Bow tie decomposition of dark web.

E. BOW TIE DECOMPOSITION

The bow-tie structure of the World Wide Web as proposed by Broder *et al.*, 2000 [2] consists of six mutually disjoint sets of nodes called components which are described as follows:

LSCC: The Largest Strongly Connected Component of the graph and is also called CORE.

IN: The set of nodes excluding those in LSCC and are reachable to CORE.

OUT: The set of nodes excluding those in LSCC and are reachable from CORE.

TUBES: The set of nodes excluding those in LSCC, IN and OUT such that they lie in between the directed path from IN to OUT.

TENDRILS: The set of nodes excluding all the above-listed nodes such that they are reachable from IN or can reach to OUT.

DISCONNECTED: The set of all the remaining nodes.

To examine the presence of the bow-tie structure in the graph, the size of its various components are computed. As already discussed in the preceding sub-section, S1 is the LSCC having 1796 nodes; all the other components are calculated using LSCC. Fig. 8 shows the results obtained. It can be seen that the OUT is much bigger as compared to the IN. The big size of OUT is because there are many of the largest out-degree nodes in CORE that connect to the majority of non CORE nodes. All the nodes of DISCONNECTED are the isolated nodes. The INTENDRILS is the subset of TENDRILS having nodes that are reachable from IN and OUTTENDRILS contains nodes from TENDRILS that can reach to OUT. Of 22.73% of TENDRILS, 20.15% belongs to INTENDRILS and the remaining 2.58% is OUTTENDRILS.

A comparison of the size of the bow tie components of the Tor web graph with the size of the PLD graph of the surface web by Meusel *et al.*, 2015 [4] is shown in Table 3. The size of LSCC in the dark web is comparatively small to that of the surface web. This is again attributed to the fragile connectivity of the dark web. However, the size of OUT is nearly the double of its surface counterpart. The DISCONNECTED

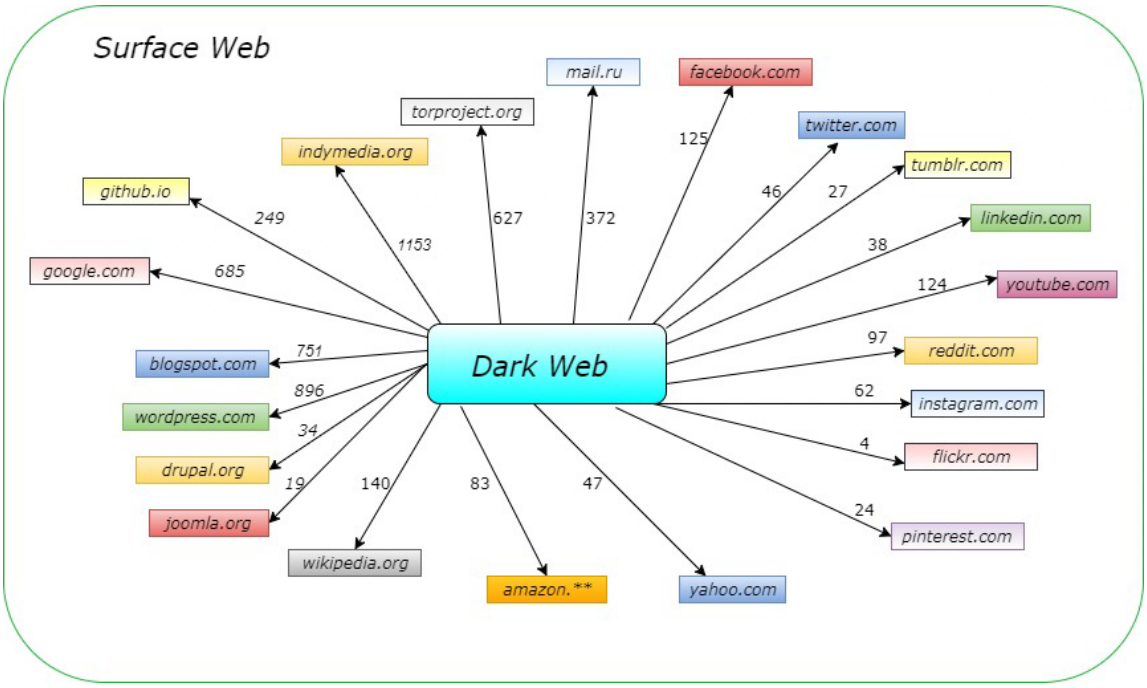

FIGURE 9. Pictorial representation of connectivity of dark web to the surface web.

TABLE 3. comparison of size (in % of nodes) of bow tie components of dark web and surface web.

Bow Tie Component	Dark Web	Surface Web
LSCC	3.73	51.94
IN	1.8	7.65
OUT	59.71	30.98
TUBES	5.79	0.04
TENDRILS	22.73	1.2
DISCONNECTED	6.24	8.2

are almost similar in size. One thing common between dark and surface is the relative size of OUT, which contains a significant number of vertices of the network than the other components of the bow tie. Overall, the Tor network does not have a typical bow-tie structure because the four of its components (LSCC, IN, TUBES and TENDRILS) are either very small or very large as compared to their surface web counterparts.

F. CONNECTIVITY TO THE SURFACE WEB

A total of 19386 outgoing links were obtained from Tor dark web towards surface web and of which, 5406 links were representing URLs of unique websites. Besides these, another 954 URLs were found ending with suffix *.i2p* that represents dark web sites running through the I2P network.[2] Fig. 9 shows the pictorial representation of the data collected. For the sake of simplicity, only the domains having an in-degree greater than ten are shown. Also, the domains which are commonly known in the surface web are shown. All the remaining domains are not included. Note that the weight on the edges denotes the in-degree of the websites from the dark web and is a sum of all the links to each of the sub-domains of that particular website.

indymedia.org is the top surface web site having 1153 incoming links from the dark web. As the dark web platform is readily used to disseminate uncensored news by journalists and whistleblowers [28], hence it can be inferred that the popularity of indymedia.org in the dark web is driven by the services it provides to journalists and whistleblowers.

The number of links to content management and blogging services like *WordPress, Joomla, Blogspot* indicates that the Tor hidden services mainly used pre-defined templates to post their content and share their thoughts and expression. Furthermore, there are many links to most of the commonly accessed social networks on the surface web. There were 124 incoming hyperlinks from the dark web pointing to the adult content websites on the surface web.

LINKAGE TO TOP LEVEL DOMAIN (TLD) IN THE SURFACE WEB

To identify the leading Top Level Domains (TLD) having maximum links from the dark web, the following 11 suffixes have been selected and the number of links from the dark web to each of the suffixes is obtained. The remaining suffixes were put under a group called *Others*. Table 4 shows the information thus retrieved.

The selected top 11 suffix contributes to ∼63% of total links with *.com* and *.org* together have close to half of the total links. Coming to country-wise TLD, *.de* (Germany) occupies the first position followed by *.ru* (Russia) and *.fr* (France). This finding may be related to the study [5] that German, Russian and French are the top non-English language with Tor hidden services. Out of total *.ru* domains, 372 domains

[2]https://geti2p.net

TABLE 4. TLD wise distribution of incoming links from dark web.

TLD	Count
com	3982
org	3676
net	1584
de	1247
ru	816
fr	361
io	165
info	153
uk	88
eu	102
edu	96
Others	7116

were of mail *.ru* an Internet giant of Russia which has a reach to around 86% of Internet users in the country [29]. One interesting fact that can be observed is the significant number of links to *.edu* domains because the dark web has a bad image of hosting unethical content [14], [30]. These *.edu* links are of the institutes like Harvard, Stanford, Princeton, MIT, etc. On closer inspection, it was found that most of these *.edu* links were coming from a single Tor hidden service that provides academic and research papers for free that otherwise need a subscription for access.

VI. DISCUSSION

The low in-degree value of the majority of the nodes suggests that the hidden services did not advertise themselves much. These hidden services may have dedicated and trusted users who very well know their address thereby eliminating the need to broadcast their URL address. Also, the lifecycle of the hidden services is very short [7]; therefore, they may not care about advertising themselves. By not revealing the URLs, the hidden services may keep the law enforcement agencies at bay to some extent. However, the question now arises is how trusted users know about their favorite hidden services? Maybe they have platforms or discussion forums on the surface web where they are in constant touch with the owner of the hidden services. In fact, many of the hidden services do have connections to some of the most popular social networks and micro-blogging platforms on the surface web. For example, a hidden service promoting terror propaganda on the dark web may have online social networks on the surface web that secretly inform their users regarding their location on the dark web. The same theory may also be applied to the isolated nodes. However, further research is required in this direction to confirm this assumption.

A significant chunk of the hidden services did not prefer linking to the other services as evident from Fig. 3. They are not interested in providing out-going links to other services. The reason for this deviant behavior may again be caused by their brief lifespan [7]. The high churn in the dark web ecosystem does not guarantee that the out-going link will always end at up and running service. Also, the dark web environment is competitive for doing business [31] and is always vulnerable to law enforcement actions. Any out-going link from the service would increase the chance of a user to navigate to all the other services.

The Tor network is resilient to the removal of the crucial nodes from the graph structure. Nearly fifty percent of the nodes remain intact in the largest connected component even after the elimination of the highest out-degree hubs. Thus the Tor network may not be much affected by law enforcement and government actions to disrupt the network. Moreover, the shutdown of even a single node by law enforcement agencies is a costly and ineffective measure [32]. However, the highest out-degree hubs may control the movement of new users across several hidden services in the largest connected components. When such hubs (like Wikis/Directories) which act as an intermediate relay to the rest of the hidden services in the connected component, are taken down by law enforcement agencies, it may stop many first-time users from accessing Tor. Old users may still be able to access the connected component if they have access to any of the other nodes.

The introvert nature of the hidden services helps them to sustain a strongly connected component of a similar type. The average distance between any connected pair is also low. The law enforcement organization may take advantage of this nature to shut down all the services simultaneously. For example, all the clones and the related forums of a marketplace can be traced by following successive links in the strongly connected component. This would disrupt the user base associated with the marketplace. The top five out-degree nodes provide links to around 93% of the nodes in the graph. In terms of law enforcement agencies, these top out-degree hub could act as a gateway to the entire dark web network for monitoring and further investigation.

The presence of out-going links to the I2P network may suggest the expansionist policy of the owner of hidden services to the other dark web networks. This may be beneficial in two ways; firstly, it may increase their chance of getting more users. Secondly, in case of the shutdown on one network, they can instantly shift their operations to the other networks.

The out-degree distribution shows that only a few nodes in the Tor network contain a vast number of out-going links while all other nodes have minimal connections. In the above discussion, we also noticed that the Tor network is tolerant to the elimination of important hidden services. In light of these findings, we can say that the Tor network exhibits the characteristics of a scale-free network [33]. This finding drives us to check whether the Tor network also exhibits the small-world network properties. In a small-world network, the two nodes are separated by the average distance of length six or less [34]. The average distance in the Tor network was 4.32 indicating it of being a small-world type. Thus we can say that the Tor network despite differing in various aspects inherently share the small-world and scale-free properties found in the surface web [18], [35].

Every study has some limitations which need to be discussed. The data analyzed in this paper was collected from the

seeds and not from the random samples which may lead the crawler to extract new links from the connected components of the respective seeds. Thus the dataset and the associated results may represent the seed hidden services. For example, if the majority of seeds were related to Bitcoin then it may be possible that the fresh links are also related to Bitcoin. However, the network structure remains unaffected by the content of the seeds.

As discussed, many hidden services would keen to remain invisible by reducing their incoming hyperlinks, therefore only an exhaustive brute force scraping of new hyperlinks could possibly explore each of the Tor hidden services. However, our crawler terminates after three successive rounds of scraping the freshly discovered hyperlinks. Consequently, the collected dataset contains a relatively small number of hidden services as compared to the number claimed by the Tor Project Inc.

Finally, we have analyzed the web graph only at the domain level. The analysis at the sub-domain and individual page level uncover several characteristics of the web graph that may differ from the one analyzed in the current work.

VII. CONCLUSION

In this study, the authors have performed an analysis of the Tor web graph at the domain level. The data has been obtained by running the web crawler tailored in Python to specifically cover the Tor dark web. The graph constructed from the collected data consists of 48,174 nodes and 1,03,526 edges where a node represents a Tor hidden service and the hyperlink between two hidden services is represented by an edge. The in-degree distribution of the web graph seems to follow the power-law distribution as the log-log plot of in-degree distribution do resemble roughly a straight line. Out-degree distribution does not follow a power law. Also, the PageRank distribution of the nodes was obtained and it was found that the PageRank and in-degree of a node are highly correlated. On the other hand, the out-degree of a node was found to be highly correlated to its eigenvector centrality.

On comparing the connectivity of the dark web, it was found that the average distance between connected pairs is almost equal to that on the surface web. However, the graph was disconnected as a whole. The bow-tie structure of the dark web was dissimilar from that of the surface web in the sense that the IN and LSCC were much smaller in size and OUT was significantly bigger. Despite poor connectivity in the graph, all the bow tie components were present though smaller in size.

Finally, the study discovers the links to the regular Internet from the dark web. A large number of such links were found, which were connecting to many of the popular surface websites like Facebook, Google, Amazon, etc. The majority of the links were to social networks, web content management, news and adult content. The TLD *.com* and *.org* were predominantly present while *.de* and *.ru* were among the top in country wise TLD.

A deeper study of the problem with a relatively large sample size is needed to obtain the properties of the Tor web graph with greater details. Also, there is a need to expand this study to cover other dark web networks which shall bring out a more clear picture of these encrypted networks.

ACKNOWLEDGMENT

This research was supported by Taif University Researchers Supporting Project number (TURSP-2020/231), Taif University, Taif, Saudi Arabia.

REFERENCES

[1] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "The Web as a graph: Measurements, models, and methods," in *Computing and Combinatorics* (Lecture Notes in Computer Science), T. Asano, H. Imai, D. T. Lee, S. Nakano, and T. Tokuyama, Eds. Berlin, Germany: Springer, vol. 1627, 1999, pp. 1–17.

[2] A. Broder, R. Kumar, F. Maghoul, and P. Raghavan, "Graph structure in the Web," *Comput. Netw.*, vol. 33, nos. 1–6, pp. 309–320, Jun. 2000.

[3] O. Lehmberg, R. Meusel, and C. Bizer, "Graph structure in the Web: Aggregated by pay-level domain," in *Proc. ACM Conf. Web Sci.*, Bloomington, IN, USA, 2014, pp. 119–128.

[4] R. Meusel, "The graph structure in the Web–analyzed on different aggregation levels," *J. Web Sci.*, vol. 1, no. 1, pp. 33–47, Aug. 2015.

[5] M. W. Al Nabki, E. Fidalgo, E. Alegre, and I. de Paz, "Classifying illegal activities on tor network based on Web textual contents," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 35–43.

[6] E. Jardine, "The dark Web dilemma: Tor, anonymity and online policing," *SSRN Electron. J.*, vol. 21, pp. 1–24, Dec. 2015. [Online]. Available: https://www.cigionline.org/sites/default/files/no.21.pdf

[7] G. Owenson, S. Cortes, and A. Lewman, "The Darknet's smaller than we thought: The life cycle of Tor hidden services," *Digit. Invest.*, vol. 27, pp. 17–22, 2018, doi: 10.1016/j.diin.2018.09.005.

[8] P. F. Syverson, D. M. Goldschlag, and M. G. Reed, "Anonymous connections and onion routing," in *Proc. IEEE Symp. Secur. Privacy*, Oakland, CA, USA, Dec. 1997, pp. 44–54.

[9] E. Jardine, "Tor, what is it good for? Political repression and the use of online anonymity-granting technologies," *New Media Soc.*, vol. 20, no. 2, pp. 435–452, Feb. 2018, doi: 10.1177/1461444816639976.

[10] M. Chertoff and T. Simon, "The impact of the dark Web on Internet governance and cyber security," *Global Commission Internet Governance*, vol. 6, pp. 1–18, May 2015. [Online]. Available: https://www.cigionline.org/sites/default/files/gcig_paper_no6.pdf

[11] E. Jardine, "Privacy, censorship, data breaches and Internet freedom: The drivers of support and opposition to dark Web technologies," *New Media Soc.*, vol. 20, no. 8, pp. 2824–2843, Aug. 2018, doi: 10.1177/1461444817733134.

[12] K. Finklea and C. A. Theohary, "Cybercrime: Conceptual issues for Congress and U.S. law enforcement," Congressional Research Service, Washington, DC, USA, Tech. Rep. R42547, 2015.

[13] M. Faizan, R. A. Khan, and A. Agrawal, "Ranking potentially harmful Tor hidden services: Illicit drugs perspective," *Appl. Comput. Inform.*, vol. 2020, pp. 1–12, Jul. 2020. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1016/j.aci.2020.02.003/full/html, doi: 10.1016/j.aci.2020.02.003.

[14] C. Guitton, "A review of the available content on Tor hidden services: The case against further development," *Comput. Hum. Behav.*, vol. 29, no. 6, pp. 2805–2815, Nov. 2013, doi: 10.1016/j.chb.2013.07.031.

[15] K. Soska and N. Christin, "Measuring the longitudinal evolution of the online anonymous marketplace ecosystem," in *Proc. 24th USENIX Secur. Symp.*, 2015, pp. 33–48.

[16] M. Bernaschi, A. Celestini, S. Guarino, and F. Lombardi, "Exploring and analyzing the tor hidden services graph," *ACM Trans. Web*, vol. 11, no. 4, pp. 1–26, Sep. 2017.

[17] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the Web for emerging cyber-communities," *Comput. Netw.*, vol. 31, nos. 11–16, pp. 1481–1493, May 1999.

[18] A.-L. Barabási, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: The topology of the world-wide Web," *Phys. A, Stat. Mech. Appl.*, vol. 281, nos. 1–4, pp. 69–77, Jun. 2000.

[19] D. Donato, L. Laura, S. Leonardi, and S. Millozzi, "The Web as a graph: How far we are," *ACM Trans. Internet Technol.*, vol. 7, no. 1, pp. 1–23, 2007.
[20] M. Serrano, A. Maguitman, M. Boguñá, S. Fortunato, and A. Vespignani, "Decoding the structure of the WWW: A comparative analysis of Web crawls," *ACM Trans. Web*, vol. 1, no. 2, pp. 1–25, 2007.
[21] (2020). *SOCKS*. [Online]. Available: https://en.wikipedia.org/wiki/SOCKS/
[22] (2020). *The Hidden Wiki*. [Online]. Available: https://thehiddenwiki.org/
[23] (2020). *TorChat*. [Online]. Available: https://en.wikipedia.org/wiki/TorChat/
[24] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.
[25] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, Nov. 2009.
[26] L. Page, S. Brin, R. Motwani, and T. Winograd. (1999). *The PageRank Citation Ranking: Bringing Order to the Web*. [Online]. Available: http://ilpubs.stanford.edu:8090/422/
[27] A. Sarkhel. (2016). *The Deep, Dark Side of Web! How People Are Getting Drugs, Guns Delivered at Doorstep*. [Online]. Available: https://economictimes.indiatimes.com/tech/internet/the-deep-dark-side-of-web-how-people-are-getting-drugs-guns-delivered-at-doorstep/articleshow/53407720.cms/.
[28] M. Chertoff, "A public policy perspective of the Dark Web," *J. Cyber Policy*, vol. 2, no. 1, pp. 26–38, 2017.
[29] (2020). *Mail*. [Online]. Available: https://en.wikipedia.org/wiki/Mail.Ru/
[30] A. Biryukov, I. Pustogarov, F. Thill, and R.-P. Weinmann, "Content and popularity analysis of tor hidden services," in *Proc. IEEE 34th Int. Conf. Distrib. Comput. Syst. Workshops*, Jun. 2014, pp. 188–193.
[31] M. Paquet-Clouston, D. Décary-Hétu, and C. Morselli, "Assessing market competition and vendors' size and scope on AlphaBay," *Int. J. Drug Policy*, vol. 54, pp. 87–98, Apr. 2018, doi: 10.1016/j.drugpo.2018.01.003.
[32] D. Décary-Hétu and L. Giommoni, "Do police crackdowns disrupt drug cryptomarkets? A longitudinal analysis of the effects of operation onymous," *Crime, Law Social Change*, vol. 67, no. 1, pp. 55–75, Feb. 2017.
[33] A.-L. Barabási and E. Bonabeau, "Scale-free networks," *Sci. Amer.*, vol. 288, no. 5, pp. 60–69, 2003.
[34] S. Milgram, "The small world problem," *Psychol. Today*, vol. 1, no. 1, pp. 61–67, May 1967.
[35] A. Barab si, "The physics of the Web," *Phys. World*, vol. 14, no. 7, pp. 33–38, 2001.
[36] C. N. Ziegler, "Fully-automatic Web data extraction," in *Encyclopedia of Database Systems*. Boston, MA, USA, 2018, doi: 10.1007/978-0-387-39940-9_1159.
[37] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, "Web data extraction, applications and techniques: A survey," *Knowl.-Based Syst.*, vol. 70, pp. 301–323, Nov. 2014.
[38] S. Li, C. Chen, K. Luo, and B. Song, "Review of deep Web data extraction," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2019, pp. 1068–1070.
[39] B. Potvin and R. Villemaire, "Robust Web data extraction based on unsupervised visual validation," in *Intelligent Information and Database Systems* (Lecture Notes in Computer Science), vol. 11431, N. Nguyen, F. Gaol, T. Hong, and B. Trawiski, Eds. Cham, Switzerland: Springer, 2019, pp. 77–89.

ABDULLAH ALHARBI received the Ph.D. degree from the University of Technology Sydney, Australia. He is currently an Assistant Professor with the Department of Information Technology, Taif University. His research interests include human–computer interaction, information systems modeling, streaming data, cybersecurity, big data analytics, and data science.

MOHD FAIZAN received the bachelor's degree from the National Post Graduate College (University of Lucknow), India, in 2013, and the master's degree in information technology, in 2015. He is currently pursuing the Ph.D. degree in information technology from Babasaheb Bhimrao Ambedkar University (A Central University), India. He has more than two years of research and teaching experience. He has published and presented articles in refereed journals and conferences. His research interest includes security engineering.

WAEL ALOSAIMI was born in Saudi Arabia, in 1979. He received the B.Sc. degree in computer engineering from King Abdulaziz University, in 2002, and the M.Sc. degree in computer systems security and the Ph.D. degree in cloud security from the University of South Wales, in 2011 and November 2016, respectively. From 2002 to 2004, he worked with Saline Water Conversion Corporation (SWCC), as an Instrument and Control Engineer. Then, he served as a Trainer for the Technical and Vocational Training Corporation (TVTC), till 2008. Then, he joined Taif University as a Teaching Assistant. It provides him with a scholarship to pursue his studies in the U.K. Since 2017, he has been an Assistant Professor with the Department of Computer Engineering, Taif University. He has many publications in peer-reviewed conferences and journals. His current research interests include cloud computing, cloud security, information security, network security, E-health security, the Internet of Things security, and data science.

HASHEM ALYAMI received the bachelor's degree in computer Science from Taif University, Saudi Arabia, in 2007, the master's degree in secure computer system from the University of Hertfordshire, U.K., and the Ph.D. degree from the University of Reading, U.K. He is currently an Assistant Professor with the Collage of Computer and Information Technology, Taif University. His research interests include cyber-security, artificial intelligent, and data science.

ALKA AGRAWAL received the Ph.D. degree from Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow. She is currently working as an Assistant Professor with Babasaheb Bhimrao Ambedkar University. She is also a Passionate Researcher. She has research/teaching experience of more than 12 years. She is working in the fields of big data security, genetic algorithms, and software security. She has published a number of research articles in national and international journals. Her research interests include software security and software vulnerability.

RAJEEV KUMAR received the master's and Ph.D. degrees in information technology from Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, India, in 2014 and 2019, respectively. He is currently working as an Assistant Professor with the Department of Computer Application, Shri Ramswaroop Memorial University, Lucknow, and a Guest Faculty with the Department of Information Technology, Babasaheb Bhimrao Ambedkar University (A Central University). He has more than five years of research and teaching experience. He is a young and energetic Researcher and holds two major projects (With PI) funded by the University Grants Commission, New Delhi, and the Council of Science and Technology, Uttar Pradesh (CST-UP), India. He has also published and presented articles in refereed journals and conferences. His research interest includes security engineering.

RAEES AHMAD KHAN (Member, IEEE) is currently working as a Professor and Head of the Department of Information Technology Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, India. He has more than 20 years of teaching and research experience. He has published more than 300 research publications with good impact factors in reputed international journals and conferences, including IEEE, Springer, Elsevier, Inderscience, Hindawi, and IGI Global. He has published a number of National and International Books (authored and edited) (including Chinese Language). His research interests include security engineering and computational techniques.