Assessing the Health of the Dark Web:
An Analysis of Dark Web Open Source Software Projects

Samuel Onyango, Emilie Steenvoorden, Joram Scholten, and Slinger Jansen(✉)

Department of Computer Science, Utrecht University, Utrecht, The Netherlands
{s.o.onyango,e.r.m.steenvoorden,e.j.scholten2}@students.uu.nl,
slinger.jansen@uu.nl

Abstract. A hidden part of the World Wide Web is known as the Dark Web, featuring websites that cannot be indexed by traditional search engines. Many open source software products are used to access and navigate through the Dark Web. Together they form the Dark Web open source software ecosystem. Research on this ecosystem is scarce and research on the ecosystem health is non-existent, even though ecosystem health is an useful indicator of the livelihood of an ecosystem. The goal of this research is to evaluate the health of the ecosystem through an assessment of Tor, I2P and GitHub. The Open Source Ecosystem Health Operationalization framework is used to help perform this assessment. Eight metrics from the framework are selected, which are measured using the data collected. Analysis of Tor and I2P metrics suggest that there has been an increase in Tor and I2P user activity in the recent past. Added knowledge, spin offs and forks and usage indicate active participation and interest in Tor and I2P. There has also been an increase in the number of active GitHub Dark Web projects. However, these GitHub projects are not well-connected and only a small number of projects have a large number of contributors. There is some variety among the GitHub software projects. The framework proves to be adequately capable of determining the health of the Dark Web open source ecosystem with the available data.

Keywords: Dark web · Open source ecosystem · Software ecosystem health · Open source ecosystem health operationalization framework

1 Introduction

“Whatever is done in the dark shall be brought to light” as the saying goes. The Internet is versed and layered from the Surface Web to the Deep Web and to the Dark Web. The deeper you go, the darker it becomes. The Dark Web is known for illicit activities, including the distribution of pornography, hacking, money laundering and selling guns and drugs on marketplaces [3]. However, the Dark Web is also used by non-criminal user groups, such as political activists, whistle-blowers, journalists, law enforcement agents and the military. Benefits of the Dark Web range from operating in totalitarian regimes and protecting

P. Gregory and P. Kruchten (Eds.): XP 2021 Workshops, LNBIP 426, pp. 125–134, 2021.
https://doi.org/10.1007/978-3-030-88583-0_12

sources to corporate spying and sharing confidential information [9]. In all layers of the Internet actors are involved, who communicate, have networks and use and produce software, thus making up a software ecosystem (SECO). Open source software ecosystems (OSSECOs) are defined by Franco et al. [5] as: "A SECO placed in a heterogeneous environment, whose boundary is a set of niche players and whose keystone player is an OSS community around a set of projects in an open-common platform".

The Deep Web and particularly the Dark Web are still shrouded in mystery. Therefore, more information is needed on the state of the Dark Web and the activities that keep its "engine running". A keen interest on special OSS such as The Onion Router (Tor) and the Invisible Internet Project (I2P), that are used to access the unindexed parts of the Internet, can provide insights on activity. Due to the anonymity, information on the Dark Web users and overall traffic is scarce [13]. However, even with published metrics, measuring the Internet by accurately classifying content and traffic remains challenging.

The main goal of this work is to shed some light on the health of the dark web ecosystem, by looking at the publicly available software that targets the dark web.

1.1 Related Work

The networks Tor and I2P allow anonymous access to the Dark Web. Similar anonymization software with significantly fewer users are Freenet, JonDonym, JAP and ZeroNet. Tor, developed in 2002 by the U.S. Naval Research Laboratory, is the most widely used network to access the Dark Web [13]. Tor's aim is to create a private access to an uncensored web while respecting user's privacy by connecting them to websites through virtual tunnels.

The popularity of Tor is due to the ease of use and the reliable anonymous access [10] resulting into 2 to 2.5 million users per day [8]. The I2P network is similar to Tor and facilitates anonymous website hosting by means of encryption and relays [3]. A difference is users of I2P automatically act as a "node" to transfer information, whereas Tor users must actively decide to become a node [1]. A study in 2018 showed I2P has around 32K active users on a daily basis [6]. Note that I2P can also be used to access the Surface Web. Cilleruelo et al. [4] analyzed the connection between the Tor and I2P networks by looking at the darknet services. They found Tor and I2P operate as an ecosystem and there are clear paths between them.

In 2014, Jansen [7] introduced the OSEHO framework. Contrary to other frameworks, OSEHO offers a more granular description of metrics, divided over three pillars, that can be used to establish the health of an OSSECO. The productivity pillar deals with the ability of projects to add value within an ecosystem. Robustness illustrates how well projects of an ecosystem can cope with changes that may occur in the environment. Lastly, niche creation indicates an ecosystem's ability to reinvent itself to take advantage of new opportunities. The metrics are applied at two levels. The project level consists of analyses of projects within the SECO, while the network level puts into action different elements for this ecosystem domain. The framework is created with considerations of SECO features, including licenses, code conventions, documentation, quality and public support for the projects OSEHO is applied to various SECOs already. [2] conducted research on the ecosystem health of cryptocurrencies, by focusing on the highest-valued distinct cryptocurrencies.

1.2 Research Method

The aim of this research is to assess the health of the Dark Web OSSECO by analysing two open source software projects. Tor and I2P are selected for this research because they have the largest number of users [12]. To determine the health of the ecosystem three metrics of the OSEHO framework, robustness, productivity and niche creation, are used [7]. The OSEHO framework is suitable for this study as it majorly involves analysis of quantitative data from the Dark Web OSS project sources, which correlate well with the metrics described in the framework. On top of that, the framework is designed for OSSECOs, which makes it a good fit. Quantitative data analysis is performed to measure the metrics, as well as some descriptive data analysis to further support the measuring of the metrics. The main research question of this paper is formulated as follows: *How can open source software ecosystem health be used as a proxy for measuring activity on the Dark Web?*. To answer this question a literature review on ecosystem health, the OSEHO framework and Dark Web OSS is conducted. Thereafter, quantitative data analysis is used to identify the productivity, robustness and niche creation of the Dark Web OSSECO. Data is collected from the code hosting platform GitHub, as well as the ecosystem hubs of Tor and I2P. Tor and I2P metrics identify the amount of users of the networks over the last twenty years. GitHub data includes historical data on the number of Dark Web projects and the number of developers. GitHub Dark Web project activity refers to the number of projects that are actively being updated and maintained.

2 Applying the OSEHO Framework

This section covers the application of the OSEHO framework in five steps.

Step 1 & 2 - Set Goals and Select Ecosystem Scope
The aim of this research is to evaluate the health of the Dark Web OSSECO, by conducting an analysis on GitHub Dark Web projects, as well as Tor and I2P. First the scope of the ecosystem needs to be determined. The entire Dark Web ecosystem is too broad to capture and a single OSS project is too limited to draw meaningful conclusions about the Dark Web OSSECO. Therefore, the scope of this study is limited to OSS projects in the Dark Web ecosystem retrieved from GitHub. Data is easier to find on OSS projects due to their open nature. Identification of the projects is done using the platform GitHub, which is the most widely-used hosting service for software development projects. The focus of the research is on ecosystem activity, as it is a clear indicator of health and data on activity is openly accessible. The traffic of Tor is used as activity data

source because it is the most used network to access the Dark Web. For the OSS projects the existing networks, users, forums developers and relationships under these projects are analysed.

Step 3 - Select Metrics

The OSEHO metrics are distinctively categorized in the pillars: Productivity, Robustness and Niche creation. To adhere to the scope of this study, a selection of the most relevant metrics are shown in Table 1.

Table 1. Overview of selected OSEHO metrics and sources.

OSEHO pillar	Metrics network level	Metrics project level	Source
Productivity	Added knowledge about ecosystem (1)		Collaboration forum (StackOverflow, Reddit and phpBB)
		Spin-offs and forks (2)	GitHub
		Usage (3)	Tor and I2P metrics
Robustness	Total number of active projects (4)		GitHub
	Cohesion (5)		Programmable web (API data)
		Active contributors (6)	GitHub
		Page views and search statistics (7)	Google Trends
Niche creation	Variety in projects (8)		GitHub

The selection and exclusion criteria for the metrics rely upon available data. Besides that, understandability is a consideration. Therefore, the metrics that are less complex, but yet simple and clear are selected. Thirdly, quantitative metrics are more considered for analysis and interpretation reasons, thus eliminating more quantitative metrics. The first metric *Added knowledge about ecosystem* (1) considers the rate at which new and existing information is shared within an ecosystem. On the project level, *Spin-offs and forks* (2) are relevant as it indicates the interest of developers in the projects within the ecosystem, therefore enhancing productivity and ultimately health. *Usage* (3), as a measure of health through productivity, is taken from the end user perspective. The number of users is an indicator of how well the ecosystem is doing. Moving on to *Total number of active projects* (4). The greater the number of active projects, the higher the survival chances of an ecosystem. For this metric, data on how often projects are updated are an indicator of activity. *Cohesion* (5) is how well-connected internally and externally the project network is. The "strength in numbers" [7] relates to how a good number of *active contributors/developers* (6) within an OSS project equates to health in that ecosystem. *Page views and search statistics* (7) show the popularity of an OSSECO. The *Variety in projects* (8) metric is the only niche creation metric. Assessment of project variety can determine the number of projects collectively contributing to the extension of the

OSS projects. Manifestation of a project in various technologies creates healthy niches in the ecosystem.

Step 4 - Assess and Collect Data

This section covers an assessment of data requirements and the applied data collection techniques for all metrics. *Added knowledge about the ecosystem* (1) is measured using an analysis of collaboration forums like phpBB, Reddit and StackOverflow. Indicators of added knowledge are aggregated information, blog posts and manuals. Next, data on the metrics *Spin-off and forks* (2) are collected through GitHub. Many software projects are an extension of other projects, as mentioned in the project description on GitHub. This information is used for the spin-off metric. The number of forks per software project are indicated on GitHub as well. The *Usage* (3) metric is measured on project level. Usage is measured through activity on the platform. Tor and I2P both publish anonymous user traffic. Tor provides analysis of their users, servers, traffic, performance, onion services as well as applications on the Tor Metrics website [14]. The I2P Metrics website provides historical infrastructure data from the I2P network (https://i2p-metrics.np-tokumei.net/). *Total number of active projects* (4) metric data is obtained by analysing GitHub repositories and identifying the ones that are involved in the Dark Web OSS projects. *Cohesion* (5) data is gathered from ProgrammableWeb, which is a source of API data. The activity metrics provided by Tor and I2P are both measures of all the activity on the networks, meaning this is not limited to the activity of the Dark Web only. In 2017 it was estimated by one of the founders of Tor that the Dark Web comprises only 3% of the Tor traffic. In 2020 a research estimated that around 6.7% of Tor traffic was related to the Dark Web [8].

For *Active contributors* (6) developers on GitHub are identified who contribute to different projects related to the ecosystem. *Page views and search statistics* (7) are measured using Google trends, which gives insight on how many searches are made relating to Tor and I2P. *Variety in projects* (8) which assesses the different kinds of projects that relate to the ecosystem are identified by analysis of GitHub projects.

Many OSS projects on GitHub are intended to be used on the Dark Web. The projects that are analyzed originate from several search terms, such as “Darkweb”, “Dark Web”, “darknet”, “Tor” and “I2P”. A small amount of the search results are actually related to the Dark Web and duplicate results are filtered out. A number of GitHub projects are identified because they were linked by other Dark Web OSS projects. This data is collected manually from GitHub. In total, 260 repositories are included, with information such as the project’s name, last update date, number of favorites, contributors and forks as well as the type of software project. Through API endpoints from Stack Exchange, Reddit, ProgrammableWeb and GitHub, data is obtained. Tools like Postman are used to make HTTP requests.

Step 5 - Data Analysis and Results

Data on added knowledge about the ecosystem is presented in Table 2. Data from the forums StackOverflow (StO), Reddit (Rdt) and phBB (phpBB) show interest in Tor and I2P. In StackOverflow alone, thirty different topics on Tor are discussed provoking 31.783 responses. Data of I2P is mined from its own dedicated forum phpBB. A total of 810 questions lead to 1194 responses, showing the developers and users are actively involved in the projects.

Table 2. The productivity metrics data of Tor and I2P from different sources. The number of questions raised by users on different issues relates to these OSS projects on the forums and the number of corresponding responses. The usage represents the highest (high) and lowest (low) number of users actively involved during a particular period. The Tor data is collected as a total of users actively involved, while I2P is determined per daily usage, with high being the day with the highest recording of active users, while low representing the lowest recorded usage.

Productivity metrics	Tor		I2p	
Added knowledge	*(StO)*	*(RdT)*	*(phpBB)*	*(RdT)*
Number of questions	30	86	810	128
Number of responses	31.783	3085	1194	1367
Spin offs and forks	2590		803	
Usage	*Total users*		*Daily users*	
High	4.090.771		30.329	
Low	1.300.000		12.400	

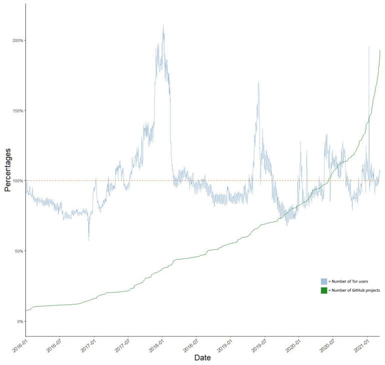

Fig. 1. The fluctuations in the number of Tor users compared to the increase of GitHub Dark Web software projects, mapped over the past five years.

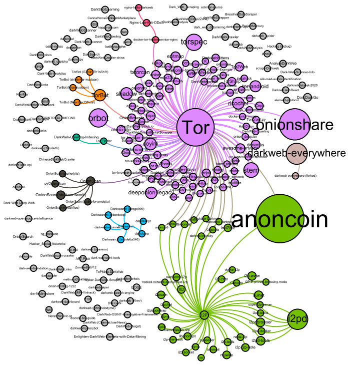

Fig. 2. Cluster visualization of the GitHub Dark Web ecosystem. The node sizes are representative for the number of active contributors a project has. Nodes with corresponding shades belong to the same cluster as assigned by the modularity algorithm.

Spin-offs and forks show many instances of projects under Tor and I2P are forked by other developers. Tor gained 2590 forks while I2P gained 803 from January 2016 to January 2021. Tor usage data shows 4.090.771 is the highest recorded number of concurrent users and 1.3 million the lowest [14]. Between 2019 and 2020 there was a sharp increase in number of users then a steady drop. Major fluctuations in these numbers are seen between 2020 and 2021 as depicted by Fig. 1. Data on I2P daily concurrent users from January 2019 till January 2021 show fluctuations with the major drops occurring between July 2019 and July 2020 with a low of 12.400 users. From November 2020 there was a steady increase to a high of over 30.000 users daily in March 2021.

With regard to robustness, roughly 260 active projects are found to be active and related to the Dark Web OSSECO. Figure 1 shows a comparison between the fluctuations in the number Tor users and active GitHub Dark Web projects. It seems like an increase in GitHub Dark Web project activity does not lead to more Tor user activity. For example, in 2021 a surge in GitHub Dark Web project activity appears, but barely any increase in Tor user activity is seen.

A Gephi network model maps the cohesion between GitHub projects in Fig. 2. It shows the projects with the largest number of active contributors, with the node size varying based on the total number of active project contributors. The model also displays the interconnectivity between projects, by looking at the edges that connect the nodes. Most of the nodes are isolated, or are connected to only one other node. The average degree of the network is 0.685, indicating a not well-connected and unhealthy network, based on the assumption that a well-connected network is healthier than networks that are not well-connected [7]. Furthermore, Tor projects appear to have far more active contributors than the I2P projects. The total number of contributors in the dataset, 1549, are not divided evenly over the projects. Out of the 260 projects that were analyzed, only 33 projects had ten or more contributors.

Google trend analysis shows Surface Web searches about Tor and I2P are significant, suggesting users are interested in the ecosystem projects Tor and I2P. From January 2016 till March 2021 major fluctuations in the number of searches are seen with the highest number being 100 and the lowest being zero.

Lastly, we analyze the variety of GitHub projects. 19 different types of software projects are identified. Most software projects are associated with crawling, networking, and security, while only a small number of projects are associated with blockchain technology, hosting, and web browsers. We conclude that there is some variety in the dark web ecosystem, but that the visible part is mostly about analyzing the dark web itself.

3 Discussion

Results show the OSSECO projects are productive, as two of the most significant contributors to health represent the productivity metric. According to [11] the productivity of OSS projects tends to be directly proportional to growth, which would validate the phenomenon observed here. Robustness is most evident through cohesion, suggesting there is some potential. Contrary to this, the number of active contributors, the page view statistics and the number of active projects weakens this pillar. Finally, niche creation is portrayed by the variety in projects. As there is some variety in the projects, we assess niche creation as moderately healthy. The results suggest an adequate amount of project variety and growth potential. Only four metrics meet the threshold required to suggest that the ecosystem is healthy, while the other four do not provide a convincing argument against a healthy ecosystem. Generally, the health of the Dark Web ecosystem is not steady at any given time. These mixed results suggest volatility, making it difficult to tell what can shift to turn the tide.

This research is subject to several validity threats. The GitHub data is collected manually, due to the choice of GitHub search terms it is possible some Dark Web OSS projects are not found. This means the used dataset could be incomplete, and should be interpreted as such. On top of that, GitHub is the only used source for OSS projects. Also, other code hosting platforms with repositories exist. The metrics that are answered using the dataset are thus only partially answered. Aside from that, Fig. 1 features Tor user data published by Tor. The metrics-timeline Git repository can be consulted to try and explain the fluctuations in the data. However, some odd peaks in the data cannot be explained by events. For example, on January 28th the user count peaked to 4.1 million, and dropped the next day to 2 million. There are no events listed in the Git repository that could possibly explain this large peak. This brings question to the legitimacy of the data. However, no good alternative for collecting data on the number of Tor users exists, due to anonymization.

3.1 Conclusion and Future Work

This paper provides an analysis of Dark Web related OSS projects determining the Dark Web OSSECO health. The OSEHO framework metrics are applied to assess the activities in this ecosystem in terms of productivity, robustness and niche creation. The study takes a unique approach by applying the framework to investigate activity on the Dark Web through the analysis of health metrics. Overall, results suggest a moderate amount of project variety and the potential to grow. All of the metrics suggest ecosystem activity but it is hard to ascertain to what percentage this represents the Dark Web as a whole. Furthermore, the health of the Dark Web ecosystem is not steady at any time.

Further research encompassing more metrics of the OSEHO framework would increase the effectiveness of the framework. The framework is implemented successfully to yield significant results showing there is activity. However, a myriad of challenges were to overcome to make this successful. First, the anonymous nature of the web made a comprehensive collection of data an uphill task leading to exclusion of some relevant metrics. Secondly, scoping the Dark Web ecosystem remains a challenge and requires future work to be done. It is ideal to have a more inclusive scope to be able to apply the framework adequately to assess Dark Web projects. Generally, the framework is practically applied to yield results that adequately answer the research question, despite the aforementioned challenges.

The OSEHO framework itself was simple to apply in practice, with many different combinations of metrics possible. Other combinations of metrics would also likely have worked, which is the real strength of the framework.

Moreover, further research is needed with the inclusion of more active OSS projects in order to measure the metrics that rely on manually collected data more accurately. This could be done by including more code hosting platforms as project sources. Additionally, more research should be done on the developers of the software project on the Dark Web to better understand who enable it.

References

1. Astolfi, F., Kroese, J. Van Oorschot, J.: I2P-the invisible internet project. Leiden University Web Technology report (2015)
2. Berkhout, M., van den Brink, F., van Zwienen, M., van Vulpen, P., Jansen, S.: Software ecosystem health of cryptocurrencies. In: Wnuk, K., Brinkkemper, S. (eds.) ICSOB 2018. LNBIP, vol. 336, pp. 27–42. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04840-2_3
3. Bradbury, D.: Unveiling the dark web. Netw. Secur. **2014**(4), 14–17 (2014)
4. Cilleruelo, C., De-Marcos, L., Junquera-Sánchez, J. Martinez-Herraiz, J.J.: Interconnection between darknets. IEEE Internet Comput. (2020)
5. Franco-Bedoya, O., Ameller, D., Costal, D., Franch, X.: Open source software ecosystems: a systematic mapping. Inf. Softw. Technol. **91**, 160–185 (2017)
6. Hoang, N.P., Kintis, P., Antonakakis, M., Polychronakis, M.: An empirical study of the I2P anonymity network and its censorship resistance. In: Proceedings of the Internet Measurement Conference 2018, pp. 379–392, October 2018
7. Jansen, S.: Measuring the health of open source software ecosystems: beyond the scope of project health. Inf. Softw. Technol. **56**(11), 1508–1519 (2014)
8. Jardine, E., Lindner, A.M., Owenson, G.: The potential harms of the Tor anonymity network cluster disproportionately in free countries. Proc. Natl. Acad. Sci. **117**(50), 31716–31721 (2020)
9. Kavallieros, D., Myttas, D., Kermitsis, E., Lissaris, E., Giataganas, G., Darra, E.: Using the dark web. In: Akhgar, B., Gercke, M., Vrochidis, S., Gibson, H. (eds.) Dark Web Investigation. Security Informatics and Law Enforcement, pp. 27–48. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-55343-2_2
10. Negi, N.: Comparison of anonymous communication networks-tor, I2P, Freenet. Int. Res. J. Eng. Technol. **4**(07), 2542–2544 (2017)
11. Scholtes, I., Mavrodiev, P., Schweitzer, F.: From Aristotle to Ringelmann: a large-scale analysis of team productivity and coordination in Open Source Software projects. Empir. Softw. Eng. **21**(2), 642–683 (2015). https://doi.org/10.1007/s10664-015-9406-4

Grey literature

12. Brown, B.: 2016 state of the dark web (2017). https://www.akamai.com/it/it/multimedia/documents/state-of-the-internet/akamai-2016-state-of-the-dark-web.pdf. Accessed 31 July 2021
13. Finklea, K.: Dark web, special report for congressional research service (2015). http://www.fas.org/sgp/crs/misc/R44101.pdf. Accessed 23 Feb 2021
14. Tor: Tor Metrics (n.d.). https://metrics.torproject.org/. Accessed 22 Mar 2021