

---

## Opensource intelligence and dark web user de-anonymisation

---

Tashi Wangchuk\*

Department of Information Technology,  
Royal University of Bhutan,  
Dewathang, Samdrup Jongkhar, Bhutan  
Email: tashiwangcukdorji@gmail.com

\*Corresponding author

Digvijaysinh Rathod

School of Cyber Security and Digital Forensics,  
National Forensic Sciences University,  
Gujarat, India  
Email: todrdigvijay@gmail.com

**Abstract:** The dark web has emerged as a platform where cybercriminals carry out illegal activities. Attempts to investigate and de-anonymise the suspicious dark web users have not been able to keep up with the pace of the dark web's flourishing coupled with dysfunctional tools and techniques. This study proposes and evaluates a dark web investigation framework using a Python-based tool to harvest data from the dark web to derive intelligence for further investigation using the available opensource intelligence (OSINT) tools. In the experimental implementation of the framework and the tool (Dark2Clear), the tool successfully scraped the hidden service URLs, harvested the e-mail addresses of the dark web users, and suspicious e-mail addresses were used as input to the OSINT tools for gathering intelligence to de-anonymise. It was observed that the framework and tool can be effectively used by the investigators to investigate and de-anonymise suspicious dark web users.

**Keywords:** hidden services; opensource intelligence; dark web; investigation framework; de-anonymisation.

**Reference** to this paper should be made as follows: Wangchuk, T. and Rathod, D. (2023) 'Opensource intelligence and dark web user de-anonymisation', *Int. J. Electronic Security and Digital Forensics*, Vol. 15, No. 2, pp.143–157.

**Biographical notes:** Tashi Wangchuk completed his BSc in Computer Science from the University of Madras and received Master's in Digital Forensics and Information Security from the National Forensic Sciences University, India. He is a Lecturer at the Department of Information Technology, Royal University of Bhutan. His research interests include

machine learning for malware detection, dark web investigations, and monitoring corporate information leakage. He is author of published research articles and independently published books.

Digvijaysinh Rathod received his BSc in Electronics and Master's in Computer Application (MCA) respectively. He completed his PhD in Computer Application from the Ganpat University (GUNI), India. He is currently working as an Associate Professor in Cyber Security in the School of Cyber Security and Digital Forensics, National Forensic Sciences University (Institution of National Importance), India. He has around 18 years of teaching and research experience, and has published more than 30 research papers in the reputed journals, conferences, and seminar and workshop proceedings. His area of interest includes mobile and web application security, blockchain and ICS/SCADA security.

---

## 1 Introduction

The portion of the World Wide Web known as the surface web or clear web which is readily accessible to normal users is very small and can be indexed by conventional search engines such as Google, Yahoo, Bing, etc. and it is accessed by the users with the help of browsers such as Chrome, Firefox, Internet Explorer, etc. The rest of the contents not indexed by the normal search engines is called the deep web (Kalpakis et al., 2016).

Chertoff and Simon (2015) stated that the dark web is a part of the deep web where the contents are intentionally hidden and not accessible to standard web browsers. The users who need anonymity take advantage and move to the dark web for conducting illegal activities ranging from trading controlled substances, fraudulent financial transactions, identity theft, etc. The portion of the internet (dark web) where illegal and criminal activities are taking place gained recognition when two students traded illegal marijuana in 1971 using the ARPANET (Adewopo, 2021), which stands for Advanced Research Projects Agency Network. The platforms for committing the crimes have increased compared to the traditional crime; now the crimes are committed on the dark web as well.

The OSINT is defined as the intelligence produced from publicly available information that is collected, exploited, and disseminated on time to an appropriate audience to address a specific intelligence requirement (Williams and Blum, 2018). Thus, OSINT can help gather business information, and identify current and future threats posed to a particular entity or organisation. As pointed out by the researcher Chad Los Schumacher, the clear web domains, e-mail addresses, payment information such as PayPal, etc. which carry a very high investigative value for successful de-anonymisation of suspicious dark web users are often found in the dark web pages due to the negligence of the users (Schumacher, 2018). Shutting down of the notorious Silk Road in 2013 by the US Federal Bureau of Investigations (Staley and Montasari, 2021) stands out as the most suitable example of the dark web users making the mistake of leaving personally identifiable information leading to their arrests and ending their illegal businesses.

Investigating and gathering the OSINT from the dark web for prosecuting the criminals and terrorists by the LEAs have gained a lot of interest but requires appropriate tools and techniques to be able to discover the dark websites providing criminal services and de-anonymising the administrators, vendors, and users (Kalpakis et al., 2016).

In the last couple of years, a significant number of criminals and terrorists have moved to the dark web, which provides anonymity to their activities and identities. On the dark web, every day, some existing websites disappear and some new ones appear. This trend is visible in the list of [.]onion sites provided to the subscribers by Hunchly in the form of a spreadsheet (Hunchly, 2021). It is observed that due to the negligence of the webmasters, vendors, and customers the dark web pages containing the clear-web-mentions such as clear-web e-mail addresses, clear-web domains, clear-web payment methods, etc. can be linked to the surface web to give valuable investigative information about the users and contribute to the success of de-anonymisation of the suspicious dark web users (Schumacher, 2018). The advancement in computing technology has brought comfort and increased the level of luxury, however, it has also brought challenges to the information security community and the law enforcement agencies (LEAs), and opportunities for criminals. To successfully investigate and de-anonymise the Dark web users, it is imperative to have an appropriate process and functional tool to gather the data from the dark web to be used for investigation. This work focused on testing and evaluating the proposed framework using a Python-based dark web scraper to harvest the clear web domains and e-mail addresses from the dark websites to investigate and de-anonymise the dark web users using OSINT tools.

## **2 Literature review**

In the following sections, the findings of various researches are presented under different thematic areas such as the underlying technological concept of surface-deep-dark web, the onion router (Tor) network and hidden services, crimes in the dark web, and harvesting data from the dark web, and dark web investigation and challenges.

### *2.1 Surface, deep, and dark web*

The internet is comprised of the surface web, deep web, and dark web. The surface web is comprised of web pages that normal search engines such as Google, Yahoo, etc. can locate and index. Whatever the search engines can index is a small part of the surface web and on the other hand, the pages that the search engines cannot locate and index is a part of the deep web (Avarikioti et al., 2018). Further, a part of the deep web contains several darknets that are aimed at providing anonymity to the users for their illegal markets of goods and services (Hurlburt, 2017).

### *2.2 Tor and hidden services*

Nath (2015) mentioned that online anonymity systems are categorised into centralised trust and distributed trust. In a centralised trust such as a virtual private network (VPN), an entity (provider of the service) would be able to know the identities of all the users and their partners, while in the distributed trust, no systems or entities would be able to

know the users and their behaviours. Such distributed trust systems are Tor, the Invisible Internet Project (I2P), and Freenet.

Tor is a free software used to protect the privacy of the users by obscuring the traffic analysis. Tor traffic is passed through multiple nodes which encrypts the information and passes it on to the next. Also, when surfing the internet using Tor, each session uses different pathways. Hosting of websites is made possible without having to disclose the location of the server and the hosted hidden services use the '[.]onion' extension, unlike the sites hosted on the clear web. Such sites are never rendered outside the Tor network (Robertson et al., 2017). Finklea (2017) mentioned that Tor refers to both the software and the network of computers that manages Tor connections.

Tor works on the concept of onion routing and the connection between the source and the destination is not made direct, instead, all the messages follow a specific pathway that is randomly calculated on the source. The pathway consists of a sequence of encrypted connections (mediating network nodes acting as the relays provided by the computers around the world signed up to participate in the network) to cover the communication tracks (Kalpakis et al., 2016). All the links encrypt and forward until the intended computer receives and decrypts the data to be displayed to the user (Collier, 2021).

The Tor also allows hosting of the contents on the servers without disclosing the location and it is known as the Tor hidden services (THSs). The Tor servers are configured in such a way that the incoming traffic is received only via the Tor network and accessed through the onion address. The Tor network interprets the '[.]onion' addresses, routes the data to and from them, and provides anonymity to the clients and the servers (Kalpakis et al., 2016).

The dark websites look no different than other clear websites, except they have a peculiar URL format that ends with '[.]onion' instead of ending with '.com', '.net', '.org', etc. That is a top-level domain suffix used by the hidden services which are reachable only via the Tor network (Razali and Suradi, 2019). The URLs are gibberish by nature and to find a link to a dark website, directories that provide a link to a specific site are used such as hidden wiki, and referral links are shared using forums and social networks. As stated by Faizan and Khan (2019), no search engine can provide a complete list of the hidden services since the normal search engines cannot index them.

### *2.3 Commerce and crimes in the dark web*

Tor is being used by researchers and journalists for censorship circumvention in their countries, the activists and journalists use it to report and communicate securely and anonymously, LEAs use it for performing undercover surveillance, and some use it to protect from criminals such as cyber-stalkers, etc. While on the other hand THSs are being used for running and staging criminal markets such as the Silk Road, viewing child exploitation materials (CEM), planning and communication of terror activities, and whistle-blowing (Nath, 2015). Although Tor provides uncensored access to the internet for communication, the Tor network hosts the infamous dark web for illegal activities such as child pornography distribution, buying and selling of illegal drugs, and sale of weapons is conducted undercover (Faizan and Khan, 2019).

The dark web offers one of the safest platforms to trade stolen and compromised data such as credit card details, social security numbers, bank accounts, social media accounts, complete individual records of persons, etc. The presence of such abundant information helps the criminals to bypass the security mechanisms put in place and the rate of crimes committed using such information is going to rise accordingly. The darknet is the root of most cybersecurity threats (Hurlburt, 2017). Business transactions in the dark web are carried out using cryptocurrencies such as Bitcoin, Ethereum, Monero, etc. A cryptocurrency enables the buyers and sellers on the dark web to conduct a transaction anonymously. Especially for buying and selling goods and services on the dark web, the escrow service model is followed (Razali and Suradi, 2019). An escrow is a financial arrangement whereby a third party holds and regulates the payment of the funds required for two parties involved in a given transaction. It helps make transactions more secure by keeping the payment in a secure escrow account which is only released when all of the terms of an agreement are met as overseen by the escrow company (Holt and Lee, 2022).

As detailed by Lee et al. (2019), transacting illegal business on the dark web using cryptocurrency involves a sequence of steps such as advertisement, discovery, negotiation, payment, and fulfilment. As per the report published by the European Monitoring Centre for Drugs and Drug Addiction in 2017, anonymisation services allow users to browse the web without having to reveal their true identity or location. Also, the anonymisation services allow the content to also be hosted anonymously by disguising the server location which is a hidden service feature. Such features have enabled the darknet marketplaces to trade illegal goods and services though the anonymisation services was not intended for illegal goods and services (Europol, 2021).

## *2.4 Harvesting data from dark web*

In a recent couple of years, a considerable number of studies have been carried out in the area of the darknet or dark web to collect posts from the forums, and gather intelligence on terrorists and their activities. There are also similar works done with varying approaches and applications. In the study carried out by the researchers Avarikioti et al. (2018) to analyse the topology and the type of contents found on the darknet, and developed a darknet spider to crawl the links recursively for 34,000 hidden services and found 10,000 hidden services online. Based on their classifier, they have observed that about half of the darknet content was related to illicit activities. The structure and the privacy analysis were carried out in 2017 by Sanchez-Rola, Balzarotti, and Santos with the help of the crawler, over 1.5 million URLs and 7,257 onion domains. Their results showed that there was a strong link between the surface web and the dark web, meaning a considerable number of websites had links, and even redirected to websites outside the [.]onion network and also it was noted that web tracking was present in the dark web (Sanchez-Rola et al., 2017).

## *2.5 Dark web investigations and challenges*

According to Koch (2019), several types of de-anonymisation efforts have been made to de-anonymise dark web users. The exploitation of the users' mistakes and human behaviour is one of the de-anonymisation techniques. A popular example is the

shutdown of the ‘Silk Road’, in 2011 created by Ross Ulbricht known as ‘Dread Pirate Roberts’, at one time in 2011 also posed as ‘altoid’ to promote his marketplace and used the same in the Bitcoin talk with his e-mail address rossulbricht@gmail.com, which enabled authorities to trace him leading to his arrests and imprisonment.

The researchers proposed and adopted an analytical framework for collecting the data from the dark web automatically using AppleScript and scraping the account details of the vendors and their marketplace listings. Their analytical framework involves the installation of Tor, creating AppleScripts for scraping, running scripts, populating the database, and identifying suspects with Maltego by successfully identifying four vendors as the most active accounts using the Maltego free tool (Hayes et al., 2018).

The dark web provides privacy and anonymity to both the users and the data unlike the clear or the surface web. It is this feature that attracts the users and facilitates the communication of secret and sensitive data for legitimate purposes for individuals such as activists, oppressed people, journalists, whistle-blowers, etc. The anonymity feature provided by the dark web also creates an ideal place for transacting illegal information, goods, services, etc. which are misused to protect individuals with criminal intentions (Kalpakis et al., 2016).

It was well noted that the dark web is used prominently for carrying out various illegal activities by criminals and there were studies conducted ranging from attempts to classify the dark websites, understand the topology and the type of content in the dark web. It was also noted that the exploitation of the users’ mistakes and human behaviour is one of the techniques for de-anonymisation, but there is no research literature focusing on this de-anonymisation technique. Therefore, the purpose of this study is to create an analytical framework for exploiting the users’ mistakes using the dark web scraper and using the OSINT tools to de-anonymise the dark web users.

### **3 Methodology**

The goal of this work was to develop a dark web scraper to collect the clear-web-mentions such as domains and e-mail addresses from the dark web pages for further investigation using the OSINT tools following the analytical framework proposed in Figure 1.

#### *3.1 Proposed framework*

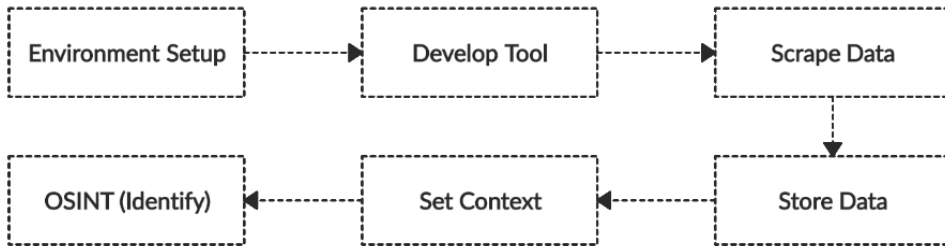
For the overall implementation and dark web investigation, the following framework was proposed and used, starting from setting up the environment for development, creating the scraper, harvesting data from the dark web, storing harvested information, setting the context to identify where the data was found and in which context was it used (illegal or legal) and then finally to perform OSINT gathering using the available tools to identify the dark web users.

#### *3.2 Collection of [.]onion seed URLs*

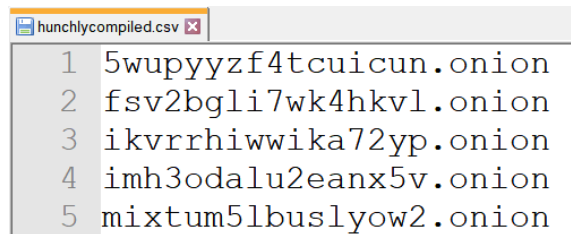
The data collection for the study was carried out in two phases, first the [.]onion URLs and then the clear web domains and e-mail addresses from the dark websites.

Hunchly has a daily hidden service report distribution provision maintained on its website (<http://www.hunch.ly/darkweb-osint/>) based on subscription. Upon subscription, Hunchly sends the identified hidden services links every day in a spreadsheet file via e-mail to the users to download and view. However, on their webpage, they caution that they do not perform the content analysis of the hidden services (Hunchly, 2021). To create a list of seed URLs for crawling and fetching more [.]onion URLs, 105,188 [.]onion links provided by Hunchly were compiled as shown in Figure 2.

**Figure 1** Framework for dark web scraping and OSINT investigation



**Figure 2** Sample [.]onion URLs compiled from Hunchly hidden services report (see online version for colours)



### 3.3 Lab setup

The Parrot operating system (OS) is a GNU/Linux distribution based on Debian. It has a full suite of tools for cybersecurity-related operations right from penetration testing, digital forensics, reverse engineering, and the environment required to develop software (Antaryami, 2021). For the development and testing of the proposed tool for dark web crawling and scrapping, the Parrot OS was used as the platform. The tools and utilities such as Python 3.8 related packages, Privoxy, and Tor bundles were installed on the base OS. The proxy (Privoxy) can be used anonymously to gain access to the Tor hidden networks (AlKhatib and Basheer, 2019). For the development and testing system to be able to access the dark web via Tor and the proxy, the Tor and the Privoxy services were kept enabled and running. The ControlPort Listener for Tor was enabled to listen on port 9051 and the new password hash was added to the `/etc/tor/torrc` file to prevent random access to the port by the agents (Mad'ar, 2021). Also, to use the Privoxy as the proxy, the Privoxy is configured to route all the traffic through the socks server at localhost port 9050.

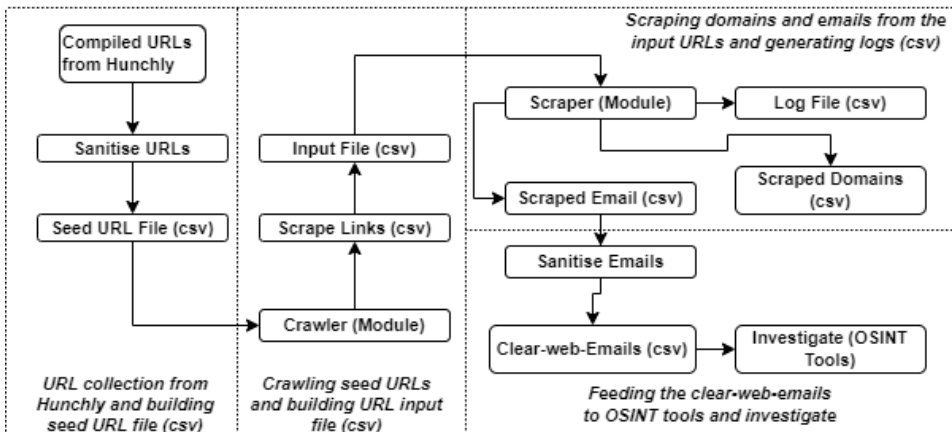
### 3.4 Scraper

The components for the developed scraper (Dark2Clear) include an interface for the user access, crawler, scraper, and sanitiser module. The user interface of the scraper tool provides the users with options in the form of a menu, such as to crawl the seed URLs, build the input.csv file from the scraped [.]onion URLs after removing the duplicate and invalid URLs, scrape the clear web domains and e-mails from the URLs provided in the input.csv file, sanitise the scraped domains and e-mails, categorise the e-mails, and managing files such as clearing the contents of the logfiles, seed files, and input files, etc. The crawler module crawls a given [.]onion URL list extracts the hidden service URLs embedded on the pages and builds the input.csv file from the scraped links respectively. The scrapper module will visit the [.]onion URLs given as the input and scrape clear web domain (e.g., example.com) and clear web e-mails addresses (e.g., username@gmail.com). The sanitiser module removes the duplicate and invalid links that were compiled from Hunchly at the first stage, then after crawling the seed URLs, the links are sanitised and the input file is being built. Later, the scraped clear web domains and e-mails are also sanitised to remove the invalid and duplicate entries, and finally, the clear-web domains and clear-web e-mails are given as output to be used as input for the OSINT tools.

### 3.5 Experimental implementation

The tool (Dark2Clear) makes the connection to dark websites hosted on the Tor network via Tor and Privoxy to ensure that security and anonymity are maintained. Figure 3 shows the overall steps and flow processes of the seed URL collection, crawling to fetch the [.]onion URLs, scraping domains and e-mails, and investigating to de-anonymise the suspicious dark web users.

**Figure 3** Overall processes involved in dark web OSINT investigation

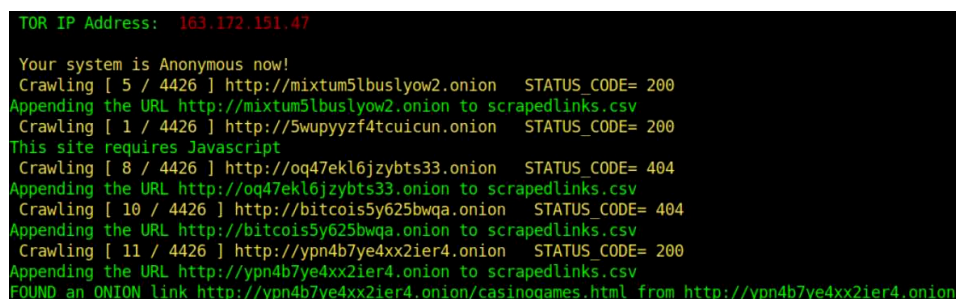




### 3.5.1 Crawling and fetching [.]onion URLs

To crawl the seed URLs and fetch the [.]onion URLs, a total of 105,188 URLs were collected and compiled from Hunchly subscriptions were curated and sanitised with the sanitiser module. After sanitising, 4,426 unique [.]onion URLs were used as the seed URL for crawling. Upon crawling using the seed URLs, a total of 73,573 links were scraped and saved to the scrapelinks.csv file, and further, it was sanitised to generate the input file (input.csv) consisting of 9,135 unique [.]onion URLs. Figure 4 is a screenshot of the tool crawling the seed URLs to fetch the [.]onion URLs.

**Figure 4** Dark2Clear tool crawling the seed URLs (see online version for colours)



```
TOR IP Address: 163.172.151.47

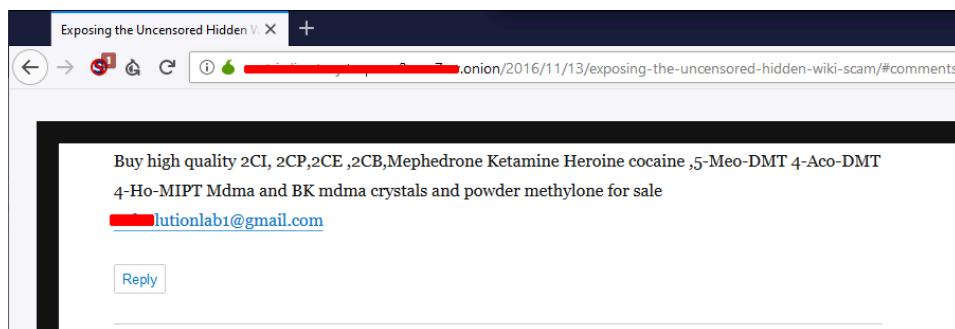
Your system is Anonymous now!
Crawling [ 5 / 4426 ] http://mixtum5lbuslyow2.onion STATUS_CODE= 200
Appending the URL http://mixtum5lbuslyow2.onion to scrapelinks.csv
Crawling [ 1 / 4426 ] http://5wupyyzf4tcuicun.onion STATUS_CODE= 200
This site requires Javascript
Crawling [ 8 / 4426 ] http://oq47ekl6jzybts33.onion STATUS_CODE= 404
Appending the URL http://oq47ekl6jzybts33.onion to scrapelinks.csv
Crawling [ 10 / 4426 ] http://bitcois5y625bwqa.onion STATUS_CODE= 404
Appending the URL http://bitcois5y625bwqa.onion to scrapelinks.csv
Crawling [ 11 / 4426 ] http://ypn4b7ye4xx2ier4.onion STATUS_CODE= 200
Appending the URL http://ypn4b7ye4xx2ier4.onion to scrapelinks.csv
FOUND an ONION link http://ypn4b7ye4xx2ier4.onion/casinogames.html from http://ypn4b7ye4xx2ier4.onion
```

### 3.5.2 Scraping clear-web mentions

The input.csv file generated after crawling and sanitising was given as the input for the scraper module to scrape the clear-web domains and the clear-web e-mail addresses having a very high investigative value for de-anonymising the dark web operators and users (Schumacher, 2018). After the clear web domains and e-mails have been scraped, it was processed through the domain and e-mail sanitiser to remove the duplicate and invalid domains, and then save the potential domains and the e-mail addresses to the separate comma separated value (CSV) files.

## 3.6 Context setting of usage

All the domains and e-mail addresses, the scraper collected cannot be treated as the starting point of investigation because the collected domains can be legitimate, genuine, and legal service providers among the domains providing illegal services. Similarly, all the collected e-mail addresses would not be abusively used on the dark web. So, it becomes very important to set a clear context as to where and how the domains and e-mails are being used. Some of the e-mail IDs collected from the dark websites were found to be used in various contexts, such as buying and selling controlled drugs, trading PayPal and Western Union accounts, etc. The redacted e-mail addresses \*\*\*\*\*lutionlab1@gmail.com, alex.\*\*\*\*\*.luna@gmail.com, mr.\*\*\*\*\*t0r@gmail.com, \*\*\*\*\*deno@gmail.com, and \*\*\*\*\*n97@gmail.com had been used in the contexts which capture the attention of the investigator. For example, the e-mail address \*\*\*\*\*lutionlab1@gmail.com was collected from a site that was suspiciously involved in the selling of controlled drugs.

**Figure 5** The context of redacted e-mail \*\*\*\*\*lutionlab1@gmail.com (see online version for colours)

### 3.7 De-anonymisation

All the e-mail addresses harvested using the scraper are not used for illegal purposes; some are used on legitimate sites while some are used for various malicious purposes. For this reason, it was important to find out the context in which the e-mail addresses were used. Till this point of implementation, the clear web e-mails collected from the dark web using the scraper were examined and their contexts of usage were identified for the selected e-mail addresses. For this sample study, the e-mail address was used as the starting point of OSINT investigation using the Maltego CE and Lampyre. The Maltego tool allows performing the link analysis after taking any combination of information such as username, phone number, e-mail addresses, IP addresses, etc. which are referred to as transforms in the context of Maltego terminology. After filling in the required information and running the transform, the Maltego tool identifies and displays wherein the web the mentioned identifiers are found (Hayes et al., 2018). The Maltego CE comes preinstalled in the Linux distributions such as Buscador Linux for investigators (Akarslan, 2021), Kali Linux, Parrot OS, etc. Lampyre is an OSINT tool that can be used to perform analysis of crimes, cyber threats, financial analytics, etc. (Daskevics and Nikiforova, 2021).

## 4 Result and discussion

The results derived after following the proposed analytical framework, the development of the Python-based tool (Dark2Clear), and the use of OSINT tools such as Maltego and Lampyre for dark web investigations are reported in the following sections.

### 4.1 Harvesting data

The scraper module successfully scraped a total of 458,470 domains and 4,068 e-mail addresses, and upon going through both script-based and manual sanitisation processes, it yielded 5,365 clear web domains and 777 e-mail addresses. Figure 6 shows a portion of the scraped e-mail addresses saved to the clear\_web\_emails.csv file.

**Figure 6** Scraped e-mail addresses saved to a file (see online version for colours)

```

sm...ow@deepdarkmail.net
Da...@gmail.com
Al...o.luna@gmail.com
Pr...ore@protonmail.com
cz...gmail.com
ha...up@protonmail.com
be...rotonmail.com
do...ecmail.pro
pe...@mail2tor.com
dw...d.biz
dw...etalks.biz
th...rgan@posteo.net
Desktop/final/clear_web_emails.csv

```

## 4.2 Generated logs

During the process of scraping the clear web domains and clear web e-mail addresses, the tool successfully maintained the records such as, from where all the domains and e-mail addresses were scraped to correlate and set the context during the investigation process later. Figure 7 shows the list of e-mail addresses (redacted) being scraped and from where the e-mail addresses are being scraped respectively. The logs are instrumental in setting the context of where and how the domains and e-mail addresses are used by suspicious users as indicated in Figure 7.

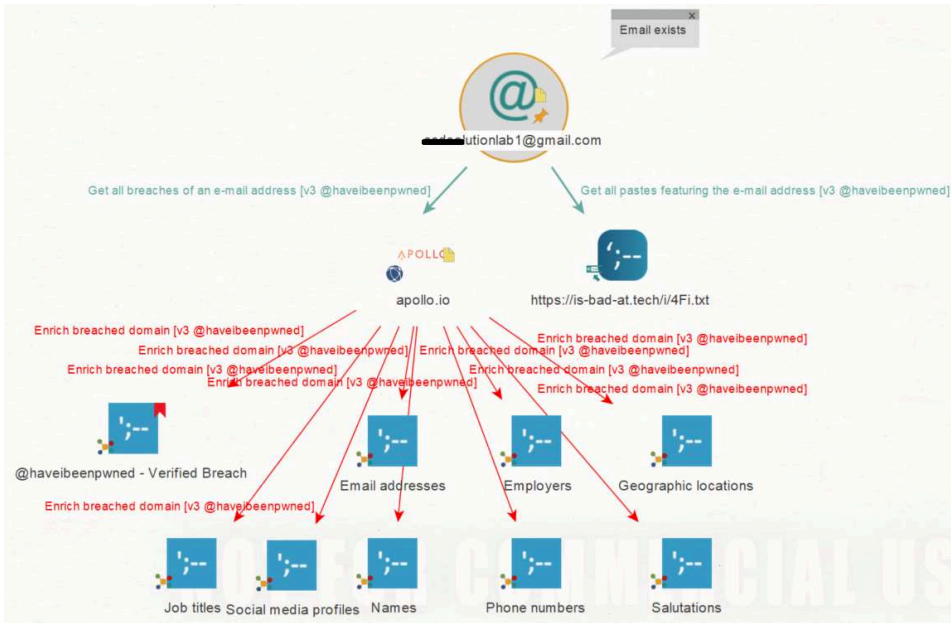
**Figure 7** Log file showing from where the e-mail addresses were collected (see online version for colours)

ID	Email Address	Source URL
3704 da	toy@protonmail.com	http://soi
3705 jili	cmail.pro	http://soi
3706 ER	ION@PROTONMAIL.COM	http://au
3707 ha	@secmail.pro	http://au
3708 da	otonmail.com	http://xil
3709 W	udm@secmail.pro	http://22
3710 toi	ibosola@yahoo.com	http://ec
3711 ad	tyet.net	http://me
3712 co	@isitwetyet.com	http://me
3713 ch	@protonmail.com	http://lyl
3714 W	udm@secmail.pro	http://22
3715 cp	secmail.pro	http://cp
3716 bit	secmail.pro	http://vq
3717 bit	secmail.pro	http://vq
3718 bit	secmail.pro	http://vq

### 4.3 OSINT with Maltego

In the attempt to gather the information based on the e-mail addresses collected from the dark web as the starting point using the Maltego e-mail transform; the following redacted e-mail addresses yielded different information related to the target as shown in Figure 8.

**Figure 8** Information gathered by Maltego for the e-mail \*\*\*\*\*lutionlab1@gmail.com (see online version for colours)

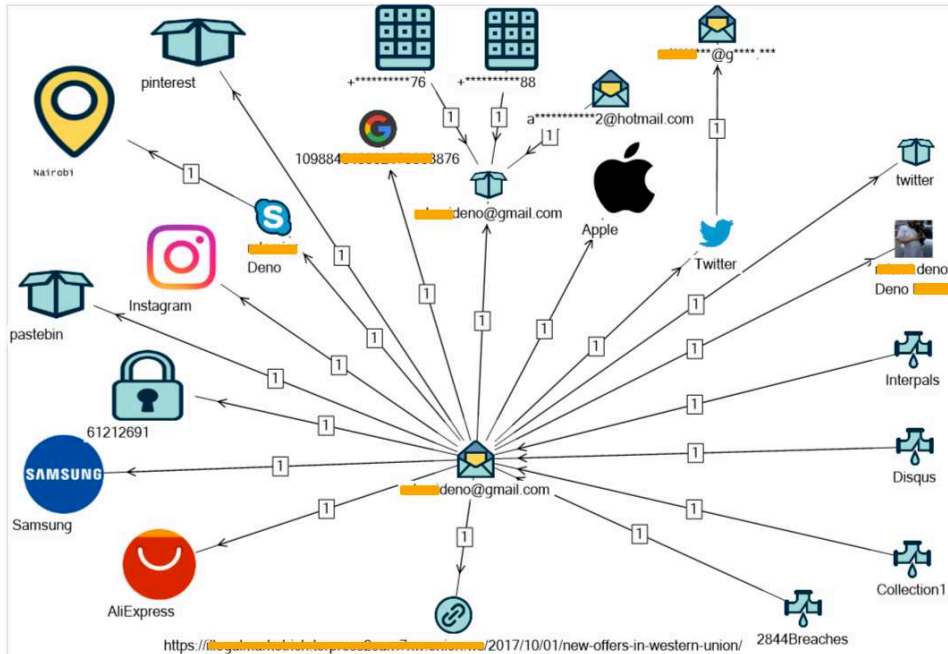


This shows the e-mail address \*\*\*\*\*lutionlab1@gmail.com exists and it was used to maintain an account on the Apollo site. However, in the data breach which involved Apollo, the details such as e-mail addresses, employers, locations, names, phone numbers, etc. were leaked. Further, the e-mail had also been found on the paste site which could have even leaked the passwords in plain text or the hashed form.

### 4.4 OSINT with Lampyre

When Lampyre was used to gather information on the target using the e-mail addresses as the starting point, the information gathered was comparatively more than that of what Maltego CE had fetched. It is because Maltego CE has limited free transforms, whereas the Lampyre seemingly provided full features but with limited search counts. Due to the limited search counts provided for the trial Lampyre account, the redacted e-mail address \*\*\*\*\*deno@gmail.com was tried. Figure 9 shows the information in the redacted form such as Skype username, Google ID, location, profile picture, etc.

**Figure 9** Information fetched by Lampyre for e-mail \*\*\*\*\*deno@gmail.com (see online version for colours)



The result of the present study demonstrates that the methodology used is suitable to perform automated crawling and scraping of personally identifiable information such as domains, e-mail addresses, etc. from the dark websites by using the Python-based script and further investigating to de-anonymise the suspicious dark web operators and users using the OSINT tools. However, some of the dark websites using JavaScript for security and anonymity reasons were not crawled or scraped. Also, sites using the robots.txt to disallow crawling were excluded from crawling and scraping. The domains and e-mail addresses were collected purely from the landing pages of the dark websites and do not include those pages protected with passwords and required user login. Although, the probable clear web domains found in the dark websites were harvested together with the e-mail addresses, for the sample OSINT investigation only the e-mail addresses were used for the demonstrative investigation.

## 5 Conclusions

From the experimental implementation and evaluation of the proposed analytical framework and the tool, it was observed that the tool effectively and efficiently crawled, and collected the domains and e-mail addresses from the dark web pages which could have high intelligence value for the dark web investigations and also the proposed framework has proven to be effective in guiding the investigator towards performing dark web investigations. The harvested data serve as the starting point (input) to the OSINT tools for performing further investigations to derive intelligence and de-anonymise suspicious and illegal dark web users. The data collected by the

tool guided by the proposed framework revealed that dark web users can knowingly or unknowingly leave personally identifiable information on the dark web which could carry a very high investigative. The outcome of this study would benefit the community of investigators either private or LEAs in de-anonymising the dark web users suspected of carrying out illegal activities.

## References

- Adewopo, V. (2021) *Exploring Open Source Intelligence for Cyber Threat Prediction*, PhD thesis, University of Cincinnati.
- Akarslan, H. (2021) 'A model proposal for analyzing open-source information in end-user computers', *International Journal of Intelligence and CounterIntelligence*, pp.1–22 [online] <https://doi.org/10.1080/08850607.2021.1899514>.
- AlKhatib, B. and Basheer, R. (2019) 'Crawling the dark web: a conceptual perspective, challenges and implementation', *J. Digit. Inf. Manag.*, Vol. 17, No. 2, p.51.
- Antaryami, A. (2021) *Comparative Analysis of Parrot, Kali Linux and Network Security Toolkit (NST)* [online] <https://doi.org/10.7939/r3-pcre-7v35> (accessed 21 December 2021).
- Avarikioti, G., Brunner, R., Kiayias, A., Wattenhofer, R. and Zindros, D. (2018) *Structure and Content of the Visible Darknet*, arXiv preprint arXiv:1811.01348.
- Chertoff, M. and Simon, T. (2015) *The Impact of the Dark Web on Internet Governance and Cyber Security*, Paper Series, No. 6, Centre for International Governance Innovation and Chatham House [online] [https://www.cigionline.org/sites/default/files/gcig\\_paper\\_no6.pdf](https://www.cigionline.org/sites/default/files/gcig_paper_no6.pdf) (accessed 24 November 2021).
- Collier, B. (2021) 'The power to structure: exploring social worlds of privacy, technology and power in the Tor project', *Information, Communication & Society*, Vol. 24, No. 12, pp.1728–1744.
- Daskevics, A. and Nikiforova, A. (2021) 'ShoBeVODSDT: Shodan and binary edge based vulnerable open data sources detection tool or what internet of things search engines know about you', in *2021 Second International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, IEEE, pp.38–45.
- Europol (2021) *Drugs and the Darknet: Perspectives for Enforcement, Research and Policy*, Europol, No. 978-92-9497-240-8 [online] [https://www.europol.europa.eu/cms/sites/default/files/documents/drugs\\_and\\_the\\_darknet\\_-\\_td0417834enn.pdf](https://www.europol.europa.eu/cms/sites/default/files/documents/drugs_and_the_darknet_-_td0417834enn.pdf) (accessed 12 November 2021).
- Faizan, M. and Khan, R.A. (2019) 'Exploring and analyzing the dark web: a new alchemy', *First Monday*, Vol. 24, No. 5 [online] <https://doi.org/10.5210/fm.v24i5.9473>, <https://journals.uic.edu/ojs/index.php/fm/article/view/9473/7794> (accessed 6 November 2021).
- Finklea, K. (2017) *Dark Web*, Congressional Research Service [online] <https://sgp.fas.org/crs/misc/R44101.pdf> (accessed 6 October 2021).
- Hayes, D.R., Cappa, F. and Cardon, J. (2018) 'A framework for more effective dark web marketplace investigations', *Information*, Vol. 9, No. 8, p.186.
- Holt, T.J. and Lee, J.R. (2022) 'A crime script model of dark web firearms purchasing', *American Journal of Criminal Justice*, pp.1–21 [online] <https://doi.org/10.1007/s12103-022-09675-8>.
- Hunchly (2021) *Hunchly – Dark Web OSINT – Free Daily Dark Web Reports* [online] <https://www.hunch.ly/darkweb-osint/> (accessed 3 October 2021).
- Hurlburt, G. (2017) 'Shining light on the dark web', *Computer*, Vol. 50, No. 4, pp.100–105.
- Kalpakis, G., Tsikrika, T., Cunningham, N., Iliou, C., Vrochidis, S., Middleton, J. and Kompatsiaris, I. (2016) 'OSINT and the dark web', in *Open Source Intelligence Investigation*, pp.111–132, Springer, Cham [online] [https://doi.org/10.1007/978-3-319-47671-1\\_8](https://doi.org/10.1007/978-3-319-47671-1_8).

- Koch, R. (2019) ‘Hidden in the shadow: the dark web – a growing risk for military operations?’, in *2019 11th International Conference on Cyber Conflict (CyCon)*, IEEE, Vol. 900, pp.1–24.
- Lee, S., Yoon, C., Kang, H., Kim, Y., Kim, Y., Han, D., Son, S. and Shin, S. (2019) ‘Cybercriminal minds: an investigative study of cryptocurrency abuses in the dark web’, in *26th Annual Network and Distributed System Security Symposium (NDSS 2019)*, Internet Society, pp.1–15.
- Mad’ar, D. (2021) *A Step-by-Step Guide How to Use Python with Tor and Privoxy*, GitHub GIST [online] <https://gist.github.com/DusanMadar/8d11026b7ce0bce6a67f7dd87b999f6b> (accessed 1 November 2021).
- Nath, C. (2015) *The Darknet and Online Anonymity*, House of Parliament: Postnote [online] <https://researchbriefings.files.parliament.uk/documents/POST-PN-488/POST-PN-488.pdf> (accessed 1 November 2021).
- Razali, N. and Suradi, N. (2019) *A Nest for Cyber Criminals: The Dark Web*, IEEE.
- Robertson, J., Diab, A., Marin, E., Nunes, E., Paliath, V., Shakarian, J. and Shakarian, P. (2017) *Darkweb Cyber Threat Intelligence Mining*, Cambridge University Press, Cambridge, DOI: 10.1017/9781316888513.
- Sanchez-Rola, I., Balzarotti, D. and Santos, I. (2017) ‘The onions have eyes: a comprehensive structure and privacy analysis of Tor hidden services’, in *Proceedings of the 26th International Conference on World Wide Web*, pp.1251–1260.
- Schumacher, C. (2018) *Investigating using Dark Web* [online] <https://www.slideshare.net/isightsoftware/investigating-using-the-dark-web> (accessed 13 November 2021).
- Staley, B. and Montasari, R. (2021) ‘A survey of challenges posed by the dark web’, in *Artificial Intelligence in Cyber Security: Impact and Implications*, pp.203–213, Springer, Cham [online] [https://doi.org/10.1007/978-3-030-88040-8\\_8](https://doi.org/10.1007/978-3-030-88040-8_8).
- Williams, H.J. and Blum, I. (2018) *Defining Second Generation Open Source Intelligence (OSINT) for the Defense Enterprise*, Technical Report, Rand Corporation.