On Strengthening SMEs and MEs Threat Intelligence and Awareness by Identifying Data Breaches, Stolen Credentials and Illegal Activities on the Dark Web

GEORGE PANTELIS, PETROS PETROU, SOPHIA KARAGIORGOU, and DIMITRIOS ALEXANDROU, UBITECH LTD, Thessalias 8 and Etolias 10, Chalandri, Athens, PC 15231, Greece

During the last decades, Dark Web content has risen in necessity in an increasingly connected world, where international anonymous networks provide access to data marketplaces and illicit multimedia material through the TOR or I2P networks. The motivation behind this paper is to gauge the current state and growth of the Dark Web in relation to the role it plays with special focus on Small and Medium-sized Enterprises (SMEs and MEs). More specifically, we devise Machine Learning and specialised Information Retrieval techniques to extract insights and investigate how the Dark Web enables cybercrime, maintains marketplaces with breached enterprise data collections and pawned email accounts. The research questions that we address concern: a) the role that the Dark Web plays for SMEs, MEs, and society in general; b) the criticality of cybercriminal activities and operations in the Dark Web exploiting threat taxonomies and scoring schemes; and c) the maturity and efficiency of technical tools and methods to curb illegal activities on the Dark Web through raising awareness via efficient text analytics, visual reporting and alerting mechanisms.

CCS Concepts: • **Information systems → Specialized information retrieval**; • **Security and privacy → Social engineering attacks**.

Additional Key Words and Phrases: Smart Threat Identification, ML/AI-based Cybersecurity, Cyber Threat Intelligence

ACM Reference Format:
George Pantelis, Petros Petrou, Sophia Karagiorgou, and Dimitrios Alexandrou. 2021. On Strengthening SMEs and MEs Threat Intelligence and Awareness by Identifying Data Breaches, Stolen Credentials and Illegal Activities on the Dark Web. In *The 16th International Conference on Availability, Reliability and Security (ARES 2021), August 17–20, 2021, Vienna, Austria.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3465481.3469201

1 INTRODUCTION

In the last two decades, illicit activities have dramatically increased in the Dark Web. Every year, Dark Web witnesses establishing new markets, in which administrators, vendors, and consumers aim to illegal data acquisition and consumption. On the other hand, this rapid growth makes it quite difficult for law and security officers to detect and investigate all these activities with manual analyses.

The detection and monitoring of Dark Web content have risen in necessity in an increasingly connected world. International networks providing illicit content have emerged in size taking advantage of the anonymity provided to them by the TOR network [4]. They use TOR and equivalent networks to hide their real Internet Protocol (IP) addresses and their physical locations. By having a framework to archive and monitor content that is available on the Dark Web serves as one part of a growing tool set to help Small and Medium-sized Enterprises (SMEs and MEs) peel back the anonymity of criminals or malicious activities against them that are operating on Dark Web marketplaces.

Manuscript submitted to ACM

The Dark Web content is not indexed by search engines like Google because the content in this layer of the web is not easily retrieved. This relies on darknets which include TOR [4] or Invisible Internet Project (I2P) [14], [22]. Dark Web sites are used for legitimate purposes as well as to conceal cybercriminal or other malicious activities [21]. In addition to these activities, there is a technology shift in organisations towards a more proactive approach to security where Situation Awareness (SA) is integrated into the Information Security Risk Management (ISRM) systems, and part of SA is the collection and processing of data which can help with managing security. This idea has given rise to practices in Cyber Threat Intelligence (CTI) which involves gathering data about exploits and cybercriminal activities, particularly in Dark Web forums [15]. Besides, cybercrime or terrorism on the Dark Web can be enacted by individuals or well-organised groups [3]. Cybercrime is increasingly accessible to anyone who wishes to engage in low-risk criminal activities while still having an impact (e.g. conducting DDoS attacks on SMEs/MEs websites is as convenient as hiring a botnet which offers DDoS-as-a-Service).

A current technical gap is that the ethical approach to collect, consume and mine all this information that is travelling in the Dark Web does not feed any third party cybersecurity service (i.e., information exchange services, sharing cyber-incidents among SMEs/MEs) to motivate decision making, improve collaboration among enterprises and take further actions. Also, all the data that are being sold in Dark Web marketplaces concern individuals or small enterprises that do not have the technical means, the skills and the financial capacity to detect their breached data, their pawned email accounts or why their web services are blocked or blacklisted.

This paper focuses on how the Dark Web is being utilised with an emphasis on cybercrime and stolen datasets, pawned email accounts, breached credentials in several marketplaces, or cyber-attack botnets available for hire, and how information retrieval and textual analytics may play the role of its adversary in order to safeguard the corporate reputation and robust email accounts for SMEs and MEs. The company sizes we are targeting while we are looking for pawned email accounts are micro (i.e., <10 employees) and small (i.e., <50 employees) enterprises. We describe these hidden spaces, shed light on the type of content they populate, the activities that they harbour including cybercrime, the nature of attention they receive, and technical approaches employed by the research community to extract insights and defeat their purpose. The proposed approach introduces a novel microservices-oriented, highly scalable and non-blocking architecture for mining the Dark Web, which benefits from Artificial Intelligence (AI) and Machine Learning (ML) to produce real-time insights to non-IT domain experts, satisfying the multi-disciplinary needs of SMEs/MEs organisations that require targeted web crawling, processing and advanced analytic services.

This work provides several contributions, the most relevant ones are summarized as follows:

- A scalable *Dark Web Crawler* along with a multi-level processing pipeline;
- Cyber-incidents evidence collection and AI-fueled intuitive textual and graph analytics;
- Multi-modal reporting and alerting mechanisms presented in visual manner; and
- Interoperability with 3rd party services, sharing capabilities and tools through Open RESTful Application Program Interfaces (APIs).

The rest of the paper is organised as follows. In Section 2, we present a literature review of the existing data retrieval and analytic methods on the Dark Web. Section 3 presents our technical approach and services developed to crawl, harmonise, curate and store content from the Dark Web sites. We also elaborate on the set of textual and graph analytics that are computed in the support of raising awareness in SMEs, MEs and society in general. Section 4 reports on the experimental and quantitative analytic results, while Section 5 concludes the paper and presents our directions for future work.

2 RELATED WORK

Many interesting works studying the topology of the TOR network, monitoring the Dark Web for cybersecurity information, identifying and predicting future cyber-incidents have appeared so far. Almost all of them highlight the existence of a dark space where people and technologies converge in order to achieve secrecy, anonymity and exchange of digital goods among the members of a community, similar to that of a physical-world's black market existing from a need for unregulated exchange of goods and services between individuals.

Bernaschi et al. [1] present an in-depth analysis of TOR web by describing three crawled datasets along with their key graph features collected over a five-month time frame. They also outline why TOR is different from the Surface Web graph and assess the relationship between contents, features and structures. The proposed approach exploits TOR features to compute textual data analytics in the support of insights extraction targeting at raising SMEs and MEs awareness about cyber-incidents they are involved.

Schäfer et al. [17] present the *BlackWidow* which relies on a Docker-based microservices architecture enabling the combination of both preexisting and customised Machine Learning tools. They represent all extracted data and the corresponding relationships mined by posts in a large knowledge graph, which is made available to security analysts for advanced search and interactive visual exploration. Compared to the proposed graph analytics, we further compute and consider among the different interconnected Uniform Resource Locators (URLs) the edge weights representing the distinct references existing between different Dark Web sites by also counting the number of cyber-attack keywords found in each site.

Koloveas et al. [8] focus on the information gathering task from Dark Web sites, forums and marketplaces, by adopting a two-phase approach to data harvesting. Initially a machine learning-based crawler is used to direct the harvesting towards websites of interest, while in the second phase state-of-the-art statistical language modelling techniques are used to represent the harvested information in a latent low-dimensional feature space and rank it based on its relevance. The proposed work eliminates its search space only in Dark Web sites, forums and marketplaces where keywords from the Malware Information Sharing Platform (MISP) [19] are found, while it also performs graph analytics in addition to textual and statistical language modelling techniques.

Narayanan et al. [12] introduce *TorBot* which is an open-source intelligence tool primarily focusing on the Dark Web sites content including their page title, description and address. It simplifies the process of identification and analysis of onion services and gathers intelligence about Dark Web by saving links to a database, retrieving emails from the sites and saving the crawled results into JSON files. What it differentiates our approach is that we further perform information retrieval and textual similarity in order to identify cyber-concepts, incidents and illegal activities within the Dark Web sites.

Samtani et al. [16] present a novel Diachronic Graph Embedding Framework (D-GEF) which operates on a Graph-of-Words (GoW) representation of hackers forum text to generate word embeddings in an unsupervised manner. They also adopt semantic displacement measures inspired by the diachronic linguistics literature in order to identify how terminology evolves over time. The introduced approach mostly considers how the content evolves or differentiates within a defined time frame in order to be used as timestamped collection of evidences or as a proof that content is changing on purpose to serve legal or ethical purposes.

Liu et al. [10] systematically identify, collect, and monitor a total of 1,212,004,819 exposed Personally Identifiable Information (PII) records across both the Dark Web and the Surface Web. Their effort resulted in identifying 5.8 million stolen Social Security Numbers (SSNs), 845,000 stolen credit/debit cards, and 1.2 billion stolen account credentials.

From the Surface Web, they identified and collected over 1.3 million PII records of the victims whose PII is exposed on the Dark Web. This is currently considered a great academic collection of exposed PII, which, has been properly anonymised. Our effort focuses on identifiable information at enterprise level, including SMEs and MEs pawned email accounts or breached data from their corporate ecosystems in order to strengthen them with the technical tools and security awareness they are currently lacking. The company sizes we are targeting while we are looking for pawned email accounts are micro (i.e., <10 employees) and small (i.e., <50 employees) enterprises.

Kobayashi et al. [7] develop an expert system with a mechanism to automate crime category classification and threat level assessment, using the information collected by crawling the Dark Web. They use a bag of words from 250 posts on the Dark Web and build an expert system which takes the frequency of terms as an input and classifies sample posts into 6 criminal categories dealing with drugs, stolen credit cards, passwords, counterfeit products, child pornography and others, hierarchically built on three (3) threat levels (high, middle, low). Our textual analytic method harvests each HTML page and computes a number vector based on the TfidfVectorizer method of nltk [11]. Nltk is an open-source Python library used for statistical natural language processing and analysis, while TfidfVectorizer is a method which transforms a given text to a vector of numbers. Each word is being replaced by a number which reflects how important the word is within the text or within a collection of texts. We then proceed by training a k-means clustering model to group together the HTML pages that have illegal or cyber-attacks related content (such as references to DDoS attacks).

In a nutshell, compared to the above mentioned approaches, the proposed microservices architecture differs by exploiting various *data analytics* and *specialised information retrieval* techniques targeting at SMEs and MEs reputation, pawned email accounts, cyber-incidents and data breached by their enterprise ecosystems. Special focus is given to optimise some Dark Web traversal and retrieval aspects and devise different algorithms in order to produce textual and graph analytics.

3 MINING THE DARK WEB TO IDENTIFY CYBER-INCIDENTS FOR SMES AND MES

In the following sections, we present the several services along with their functionalities which contribute in crawling the Dark Web, curating and mining its content. In the proposed *Threat Intelligence* framework, we used open-source frameworks.

3.1 Conceptual Architecture

In this section, we describe the conceptual architecture of the proposed *Threat Intelligence* framework, as depicted in Figure 1. The framework is built up by microservices which are presented in the following and consists of two logical modules, namely, the *Dark Web Crawler* and the *Text Analytics and Business Intelligence*. All the auxiliary services that are described support the framework's functionalities and define their integration points. We installed TOR and configure it in a way that allows us to make requests with Python to the Dark Web via a TOR's SOCKS proxy. The spider search starts from a list of websites. The crawling space is expanded by adding cyber-related URLs that the crawler finds within the retrieved sites, which lead to other sites and so on.

The Dark Web Crawler module handles the search space and processes the HTML content (e.g. CAPTCHA, login, HTML cleaning, etc.). It is also responsible for keeping the status of visited and unvisited links, store the HTML content, traversing the Dark Web and finally avoiding spider traps.

Elasticsearch [5] is responsible for the data storage, the performance of the *Threat Intelligence* framework and the preliminary analysis. Not only it supports the text indexing of the HTML content to retrieve quickly textual and temporal queries, but it also indexes the metrics of the entire pipeline, such as the number of pages fetched per minute

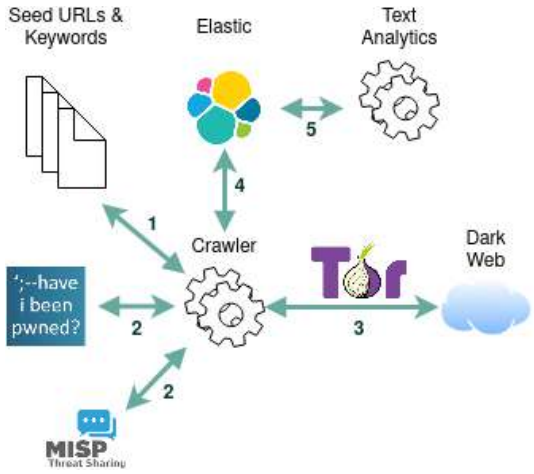

Fig. 1. Conceptual Architecture of the Threat Intelligence Framework.

or the average amount of time spent for the information retrieval. Kibana [6] is used on top of Elasticsearch for data exploration, visualisation and reporting and supports the *Text Analytics and Business Intelligence* module.

Last but not least, the purpose of the *Text Analytics and Business Intelligence* module is to extract information and knowledge relied on the collective behavior of the indexed content. This means that we perform statistical analysis for finding top-k documents based on users input (i.e., keywords). We also perform text clustering for finding similar content among Dark Web sites in order to add a layer of abstraction and group them within wider cybersecurity categories. At the end of each processing cycle, we update the document score and identify hidden connections in order to provide the next set of look up seeds.

3.2 Data Collection

Web crawlers are special applications used to create a copy of all the visited web pages for later processing. They are mainly used for indexing websites to facilitate web search engines but are also used for web archiving, web mining and web monitoring. The basic idea of web crawling is simple: given a set of starting URLs, a crawler downloads all the web pages addressed by the URLs, extracts the hyperlinks contained in the pages, and iteratively downloads the web pages addressed by these hyperlinks.

However, the last layer of the Internet is not accessible using a regular browser. URLs are continuously changing, users need to know how to find them and use a particular application or a special proxy because it exists on an alternative layer of the Surface Web. These sites are hidden behind encryption protocols such as TOR [4] or I2P [22], while this part of the Internet is called *Dark Web* [2]. Sites hidden on the TOR network are accessed via domain addresses under the top-level domain .onion. International Monetary Fund (IMF) on a report in September 2019 [9] mentioned that over 65,000 unique URLs ending with .onion exist on the TOR network. Dark Web sites are usually not crawled by generic or deep crawlers because the web servers are hidden in the TOR network and require the usage of specific protocols for being accessed. In order to crawl such sites, focused crawlers use external HTTP proxies configured to route traffic through the TOR network [23].

The data collection service of the Dark Web Crawler requires a set of initialisation steps. These include the instantiation services, the proxies and the set-up of a keywords list (which is constantly enriched) to eliminate the search space in the Dark Web to the concepts which refer to cybersecurity incidents (i.e. hacks, SQL injection, DDoS attacks, etc.), email accounts (i.e. pawned email accounts for breached corporate data, job titles, names, phone numbers, physical addresses, social media profiles, etc.), SMEs and MEs leaked corporate information and their email servers (i.e. blacklisted email servers or blocked web services). For development purposes, we chose Python 3.6 as programming language, because it offers many libraries to request, process and analyse the HTML content.

The data collection service follows an incremental pipeline to retrieve and collect the data from the Dark Web, which is concretely as follows:

- Initialise the *Dark Web Crawler* with seed URLs and user-defined keywords;
- Set the crawling depth (i.e., in our case it was set to 10) and the number of parallel non-blocking threads (i.e., in our case it was set to 5);
- Enrich user's keywords based on the MISP ontology and other publicly available ontologies;
- For each retrieved URL, we request the content from the TOR network;
- Configure a SOCKS proxy to hide/change the IP address through which we are retrieving the Dark Web content;
- Clear the cookies; and
- Change the user-agent and other request headers in order for the requested site to be unable to recognize our spider search.

The intelligence of the *Dark Web Crawler* relies on its keywords list which is continuously enriched. Based on the defined keywords list, the *Dark Web Crawler* focuses on certain content (e.g. leaked data). We utilise two sources for these keywords. These include standard keywords that we set inside the service. These keywords are further divided into categories, such as the category 'CYBER_ATTACKS' that include tokens like "ddos", "dos", "phishing", etc. At the same time, we dynamically collect keywords upon crawling. We extract keywords by using the spaCy library [18] which supports Named Entity Recognition (NER).

The processing steps, that are following, facilitate the Dark Web Crawler to massively collect data in a non-blocking and parallel manner. These include:

- Set URL seeds in a synchronised queue. This queue is used for storing the results (i.e., URLs and depth) and keep the desired order for the retrieval (i.e., through the breadth-first search algorithm);
- Retrieve the HTML page of the current URL, and perform some data curation methods to clean the text, filter out broken links and extract the important content from the page. For every HTML page, we check:
  - If it has the same domain with the parent URL (for example "myurl1.onion/posts" has the same domain with "myurl1.onion"). If it does have the same domain, we add it to the same depth with the parent URL;
  - If we have already visited this URL, we simply skip this URL;
  - If it has not been visited or has not the same domain with the parent URL, we then add it in the queue with increased depth;
  - If we reach the desired depth (i.e., in our case it was set to 10), we do not add any child links in the queue.
- Calculate a score for each HTML page that reflects the criticality or the risk of the expressed content. To calculate this score, we use the keywords and their frequency of appearance in the document (i.e., term frequency within the HTML page). Also, we use the extracted keywords with different weights, so that in the case we find keywords that are more important/interesting, the criticality score is affected more. Besides, before we search for keywords,

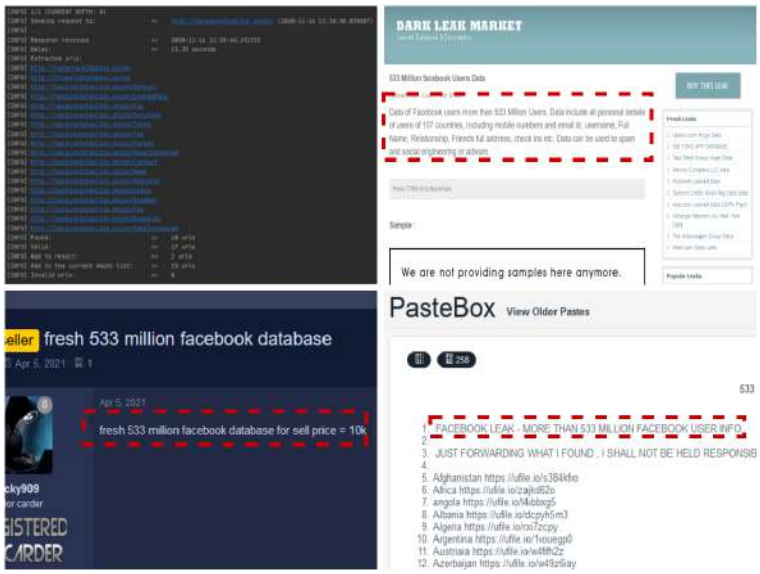

Fig. 2. Data Collection & Facebook Data Breached and Sold in the Dark Web Marketplaces.

we use a stemming technique and we convert the content in lower case letters to achieve greater match, as well as disambiguation coverage over the searched keywords. Keywords are expressed by means of N-grams.

Figure 2 depicts a running example and URLs found by the proposed *Threat Intelligence* framework. This execution is related with the Facebook data breach in April 2021[1]. During that period, we provided an extra set of keywords and N-grams to the crawler. These keywords are specific to this Facebook breach, including ('facebook', 'leak'), ('facebook', 'data'), etc. Furthermore, we provided URL seeds that are corresponding to Dark Web forums, which have been found from previous searches. We found that the Facebook data are sold in the price of 1000$, but also, we found additional data leaks including:

- Brazil's 220 million people's data[2];
- Acer.com leak[3]; and
- Covve App data leak (referred as db8151dd)[4].

3.3 Data migration, harmonisation, linkage and storage

The data migration, harmonisation and linkage services include a set of modules to store, harmonise and link the collected data with related cybersecurity concepts and ontologies. After the process of collecting the data from the Dark Web sites, we store and index them through Elasticsearch. Before we dive into more details on how we store the results in Elasticsearch, we explain how we extract and harmonise the information from the raw HTML text. When we retrieve the HTML page, we convert the raw HTML text to cleaned text that contains all the HTML content and it is used as the input for the TFIDF Vectorizer. To curate and clean the HTML text, we follow the next steps:

[1]https://www.theguardian.com/technology/2021/apr/06/facebook-breach-data-leak
[2]https://www.privacy-ticker.com/giant-database-leak-exposes-data-on-220-million-brazilians/
[3]https://www.theverge.com/2021/3/20/22341642/acer-ransomware-microsoft-exchange-revil-security
[4]https://securityboulevard.com/2020/05/covve-contacts-app-data-breach-exposes-23-million-emails-addresses-and-other-private-details/

- Use the BeautifulSoup library[5] and regular expressions to clean the HTML page;
- Remove the <script> and <style> tags along with their content;
- Remove all the formatting tags from the HTML page;
- Remove the punctuation points;
- Keep the numbers, because it is possible a keyword to have numbers such as the case of the Covve App data breach (i.e., db8151dd);
- Lower the letters of the remaining text;
- Detect the language;
- Remove the stop words (using the nltk library).

In the following, before we proceed with storing the curated content in the Elasticsearch, we make it available in a form that is easier to analyse it later. Specifically, for each HTML page retrieval, we additionally keep the following information:

- The Timestamp;
- The URL of the request;
- The root URL of the requested URL;
- The raw HTML;
- The clean/curated text of HTML;
- The critical score that we compute;
- The depth in which we reached this URL;
- The set of keywords found in this HTML page;
- The category (based on keywords); and
- The Language.

The data migration, harmonisation, linkage and storage services have been developed by using Elasticsearch which is considered a state-of-the-art Document Database and offers high performance along with a distributed architecture. Also, we use Kibana for data exploration and visualisation purposes of the computed analytics. Figure 3 depicts some indicative records and features, as we store them in Elasticsearch.

	id	url	keywords	category	critical_score
>	55	http://oniondirljacm547.onion/Cat?Id=18Page=4	btc, bitcoin, hack, hacking, password, dos, ddos, forum	COMMON, CYBER_ATTACKS	4.602
>	17,327	http://zqktlwi4fecvo6ri.onion/wiki/Denial-of-service_attack	hack, dos, ddos, games	CYBER_ATTACKS, ALIANSES	3.236
>	9,313	http://zsyvcm262oiaoc6es7bgg66xieyil6nqkh7jn5ntraghpqgudbcl3vad.onion/index.php/hire-a-hacker/	hack, hacking, password, csrf, dos, ddos, xss	COMMON, CYBER_ATTACKS	6.038
>	24,184	http://zqktlwi4fecvo6ri.onion/wiki/index.php?title=Denial-of-service_attack&action=edit	hack, dos, ddos, games	CYBER_ATTACKS, ALIANSES	3.106

Fig. 3. An Indicative Example of Collected, Curated and Stored Data in Elasticsearch from the Dark Web.

The type of information stored in Elasticsearch includes the parent URL, the keywords retrieved within the text, the identified category and the criticality score of each cyber-incident. Also, we observe that CYBER_ATTACKS category gets higher criticality score, as it is more related to cybersecurity incidents targeting at SMEs and MEs. To validate our modules, we are currently using some indicative SMEs coming from the European space. Besides, most of the

[5]https://www.crummy.com/software/BeautifulSoup/bs4/doc/

cybersecurity incidents relate with DDoS attacks, hacked passwords or pawned email accounts. The ALIANSES category relate with keywords coming with specific alias due to data breaches or equivalent incidents with the focus to identify data sold, social media data (i.e., from game accounts, social forums, etc.) or user account details sold in the Dark Web.

3.4 Text Analytics and Business Intelligence

After the data collection and harmonisation, we perform some text analytics by using state-of-the-art Machine Learning algorithms, such as clustering. The intuition is to group the HTML pages that we collect based on their content similarity. Specifically, we harvest each HTML page (after cleaning, harmonisation and linkage) and we compute a vector based on the TfidfVectorizer method of Scikit-learn [13]. We then proceed by training a k-means clustering model with 5.500 training data. We are experimenting over the clustering parameter k in order to select the best value. We observe from the top-k words of each cluster that the model clusters the HTML pages that have illegal or inappropriate content together.

With this model, we can make the crawler more focused by searching for specific patterns that we are interested in. For example, if an HTML page that we just retrieved belongs to cluster 2 (aka related with pornographic content), we stop searching because this context is irrelevant to illegal cybersecurity activities targeting at SMEs and MEs, because we are more interested in finding incidents, cyber-attacks, data breaches and malicious activities related with them, their web services, applications or their corporate email accounts. The company sizes we are targeting while we are looking for pawned email accounts are micro (i.e., <10 employees) and small (i.e., <50 employees) enterprises. The analytic results are stored to Elasticsearch in order to further feed User Interfaces (UIs) and auxiliary services including the Threat Intelligence sharing or alerting capabilities of the proposed framework. For instance, through Kibana, we provide a set of descriptive analytics by visualising the number of HTML pages found per cluster/category.

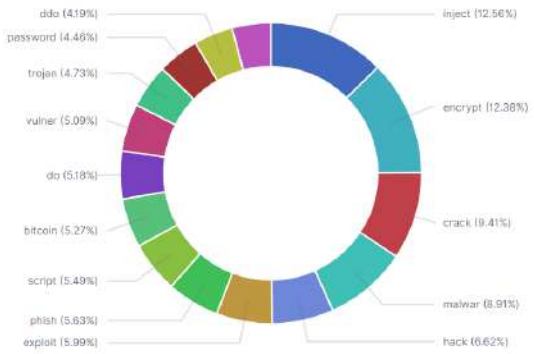

Fig. 4. Percentage of Cyber-concepts in Texts.

4 EXPERIMENTAL RESULTS

In this section, we present the results of a quantitative experimental study. After the data collection and harmonisation, we perform text analytics by using state-of-the-art Machine Learning algorithms such as clustering. The Dark Web Crawler collects more than 2.0GB (17000 HTML documents) of data per day that are further harmonised, linked, enriched and analysed. The purpose of collecting these web data sources is to extract useful information that can be

used to report about illegal trading of breached data in marketplaces, blacklisted email servers and news about the corporate rumour of small enterprises found in the Dark Web sites.

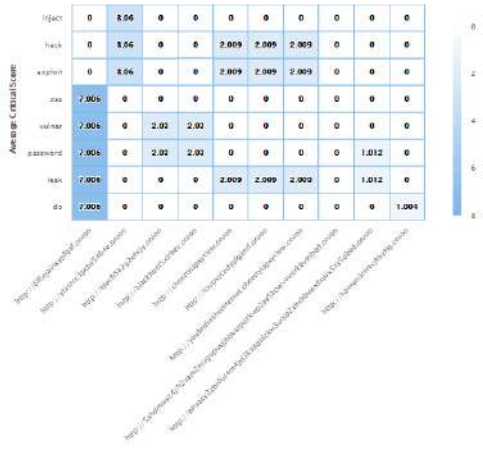

Fig. 5. Heatmap Score of URL vs Cyber-concepts.

In Figure 4, the pie indicates by means of percentage which keywords have appeared more in the Dark Web pages with content relative to cyber-attacks. For instance, hack, malware, and inject are the keywords with the greatest instances in our data collection, but at the same time they are very generic. Other documents including phising, trojan or DDoS contain much focused information about cybersecurity incidents with however fewer instances.

In Figure 5, we illustrate the root URLs with the criticality score of the keywords/cyber-concepts that are found in a heatmap. The criticality score per keyword within each URL is computed according to its category (e.g., CYBER_ATTACKS vs. ALIANSES) and the number of instances found. By using these two features, we can distinguish URLs that may contain critical information, such as breached data, preparation/call for an attack, etc. Also, we use these features to highlight and extract the most suspicious URLs, in order to create a new seed of URLs for new and more focused search.

The combination of Figures 4 & 5 give us the (root) URLs and the documents with the greatest value of cybersecurity incidents. These are serving as new seeds for a further execution of the Dark Web Crawler that contributes in going deeper in the Dark Web.

Figure 6 depicts the connections between keywords in our collection. We used statistical analysis (i.e., "significant text" query from Elasticsearch) for finding hidden connections between keywords. This means that some concepts are related in our data collection. We use sampling to find the keywords connections. For example, Figure 6 shows that "csrf" is one of several terms strongly associated with "backdoor" or "xss". It only occurs in few documents in our index and therefore most of the documents also contain "backdoor" and "xss" results. That suggests a "significant" word.

Figure 7 depicts the clustering result from *k*-means algorithm by using 5.500 training data. We are experimenting over the clustering parameter to select the best value. To reduce the time required for the data visualisation of the records kept, we use the t-SNE [20] algorithm, which is a statistical method for visualising high-dimensional data by giving each datapoint a location in a two or three-dimensional map. This algorithm results in the dimensionality reduction of the retrieved data, facilitating the observation of the graphical outcome. Based on the clustering visualisation and the top-k

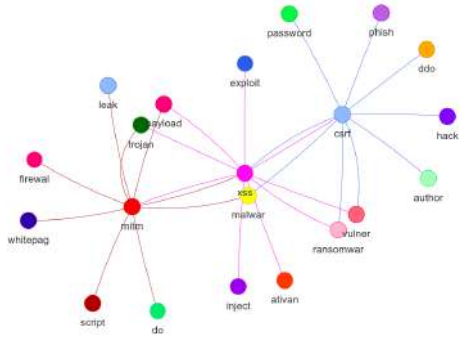

Fig. 6. Significant Text.

words of each cluster, we conclude that illegal or inappropriate content are grouped together, the cyber-incidents are grouped within the CYBER_ATTACKS category, while documents without any cyber-concept form their own cluster setting apart irrelevant concepts compared to our context.

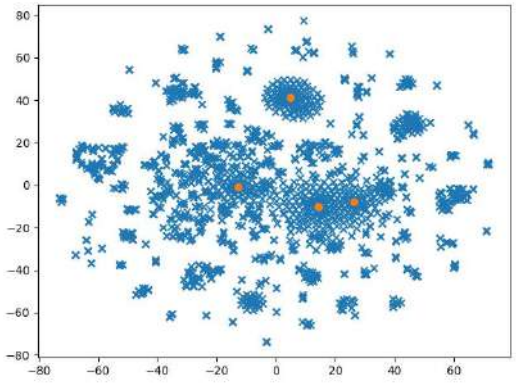

Fig. 7. Clusters of Documents.

5 CONCLUSIONS AND FUTURE WORK

This paper presented specialised information retrieval techniques on top of the Dark Web in the support of cyber-threats textual mining and business intelligence functionalities. We outlined the methods and services for data collection, harmonisation and storage, knowledge extraction, and text clustering by using open-source frameworks. We presented the functionalities of the different modules supporting the various data modalities at the Dark Web, techniques for the semantic linkage/enrichment of cyber-concepts with external taxonomies. We drilled into the details of the text analytics services that are supported for advanced SMEs and MEs cybersecurity awareness, visual reporting and alerting mechanisms for emerging cyber-incidents, events and correlation analysis based on graphs and heatmaps.

In the near future, we plan to perform a qualitative analysis and measure the accuracy of the text and graph analytics with state-of-the-art metrics, enrich the variety of analytics according to the diversity of the collected content and extend the current functionalities with more relations among the cyber-incidents.

ACKNOWLEDGMENTS

This work has received funding by the European Commission projects: a) H2020 PUZZLE - GA No 883540 (https://puzzle-h2020.com/); and b) H2020 CyberSANE - GA No 833683 (https://www.cybersane-project.eu/).

REFERENCES

[1] Massimo Bernaschi, Alessandro Celestini, Marco Cianfriglia, Stefano Guarino, Flavio Lombardi, and Enrico Mastrostefano. 2021. Onion under Microscope: An in-depth analysis of the Tor network. *arXiv preprint arXiv:2101.08194* (2021).

[2] Hsinchun Chen. 2011. *Dark web: Exploring and data mining the dark side of the web.* Vol. 30. Springer Science & Business Media.

[3] Adrian Crawley. 2016. Hiring Hackers. *Network Security* 2016, 9 (Sept. 2016), 13–15.

[4] Roger Dingledine, Nick Mathewson, and Paul Syverson. 2004. *Tor: The second-generation onion router.* Technical Report. Naval Research Lab Washington DC.

[5] Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine.* " O'Reilly Media, Inc.".

[6] Yuvraj Gupta. 2015. *Kibana essentials.* Packt Publishing Ltd.

[7] Hanae Kobayashi, Masashi Kadoguchi, S. Hayashi, A. Otsuka, and Masaki Hashimoto. 2020. An Expert System for Classifying Harmful Content on the Dark Web. *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)* (2020), 1–6.

[8] Paris Koloveas, Thanasis Chantzios, Christos Tryfonopoulos, and Spiros Skiadopoulos. 2019. A Crawler Architecture for Harvesting the Clear, Social, and Dark Web for IoT-Related Cyber-Threat Intelligence. *2019 IEEE World Congress on Services (SERVICES)* 2642-939X (2019), 3–8.

[9] Aditi Kumar and Eric Rosenbach. 2019. The Truth about the Dark Web: Intended to protect dissidents, it has also cloaked illegal activity. *Finance & Development* 56, 003 (2019).

[10] Yizhi Liu, F. Lin, Zara Ahmad-Post, MohammadReza Ebrahimi, N. Zhang, James Lee Hu, Jingyu Xin, Weifeng Li, and Hsinchun Chen. 2020. Identifying, Collecting, and Monitoring Personally Identifiable Information: From the Dark Web to the Surface Web. *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)* (2020), 1–6.

[11] Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028* (2002).

[12] P. S. Narayanan, R. Ani, and Akeem T. L. King. 2020. TorBot: Open Source Intelligence Tool for Dark Web.

[13] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[14] Regner Sabillon, Victor Cavaller, Jeimy Cano, and Jordi Serra-Ruiz. 2016. Cybercriminals, cyberattacks and cybercrime. In *2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF).* IEEE, 1–9.

[15] Sagar Samtani, Ryan Chinn, Hsinchun Chen, and Jay F Nunamaker Jr. 2017. Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence. *Journal of Management Information Systems* 34, 4 (2017), 1023–1053.

[16] Sagar Samtani, Hongyi Zhu, and Hsinchun Chen. 2020. Proactively Identifying Emerging Hacker Threats from the Dark Web. *ACM Transactions on Privacy and Security (TOPS)* 23 (2020), 1 – 33.

[17] Matthias Schäfer, Markus Fuchs, Martin Strohmeier, Markus Engel, Marc Liechti, and Vincent Lenders. 2019. BlackWidow: Monitoring the Dark Web for Cyber Security Information. *2019 11th International Conference on Cyber Conflict (CyCon)* 900 (2019), 1–21.

[18] Xavier Schmitt, Sylvain Kubler, Jérémy Robert, Mike Papadakis, and Yves LeTraon. 2019. A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS).* IEEE, 338–343.

[19] Cynthia Wagner, Alexandre Dulaunoy, Gérard Wagener, and Andras Iklody. 2016. Misp: The design and implementation of a collaborative threat intelligence sharing platform. In *Proceedings of the 2016 ACM on Workshop on Information Sharing and Collaborative Security.* 49–56.

[20] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. 2016. How to use t-SNE effectively. *Distill* 1, 10 (2016), e2.

[21] Gabriel Weimann. 2016. Going dark: Terrorism on the dark web. *Studies in Conflict & Terrorism* 39, 3 (2016), 195–206.

[22] Bassam Zantout, Ramzi Haraty, et al. 2011. I2P data communication system. In *Proceedings of ICN.* Citeseer, 401–409.

[23] Ahmed T Zulkarnine, Richard Frank, Bryan Monk, Julianna Mitchell, and Garth Davies. 2016. Surfacing collaborated networks in dark web to find illicit and criminal content. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI).* IEEE, 109–114.