Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Cybercrime threat intelligence: A systematic multi-vocal literature review

Giuseppe Cascavilla[a,*], *Damian A. Tamburri*[a], *Willem-Jan Van Den Heuvel*[b]

[a] *Eindhoven University of Technology, Jheronimus Academy of Data Science, The Netherlands*
[b] *Tilburg University, Jheronimus Academy of Data Science, The Netherlands*

ARTICLE INFO

Article history:
Received 8 October 2020
Revised 27 January 2021
Accepted 28 February 2021
Available online 5 March 2021

Keywords:
Cyber threat intelligence
Cybersecurity
Dark web
Deep web
Surface web
Topic modelling

ABSTRACT

Significant cybersecurity and threat intelligence analysts agree that online criminal activity is increasing exponentially. To offer an overview of the techniques and indicators to perform cyber crime detection by means of more complex machine- and deep-learning investigations as well as similar threat intelligence and engineering activities over multiple analysis levels (i.e., surface, deep, and darknets), we systematically analyze state of the art in such techniques. First, to aid the engineering and management of such intelligence solutions. We provide *(i)* a taxonomy of existing methods mapped to *(ii)* an overview of detectable criminal activities as well as *(iii)* an overview of the indicators and risk parameters that can be used for such detection. Second, to find the major engineering and management challenges and variables to be addressed. We apply a Topic Modelling Analysis to identify and analyze the most relevant threat concepts both in Surface and in Deep-, Dark-Web. Third, we identify gaps and challenges, defining a roadmap. *Practitioners value and conclusions.* The analysis mentioned above effectively provided a photograph of the scientific and practice gaps among the Surface Web and the Deep-, Dark-Web cybercrime and threat engineering and management. More specifically, our systematic literature review shows: *(i)* the dimensions of risk assessment techniques today available for the aforementioned areas—addressing these is vital for Law-enforcement agencies to combat cybercrime and cyber threats effectively; *(ii)* what website features should be used in order to identify a cyber threat or attack—researchers and non-governmental organizations in support of Law Enforcement Agencies (LEAs) should cover these features with appropriate technologies to aid in the investigative processes; *(iii)* what (limited) degree of anonymity is possible when crawling in Deep-, Dark-Web—researchers should strive to fill this gap with more and more advanced degrees of anonymity to grant protection to LEAs during their investigations.

* Corresponding author.
E-mail addresses: g.cascavilla@tue.nl (G. Cascavilla), d.a.tamburri@tue.nl (D.A. Tamburri), w.j.a.m.v.d.heuvel@jads.nl (W.-J. Van Den Heuvel).

https://doi.org/10.1016/j.cose.2021.102258

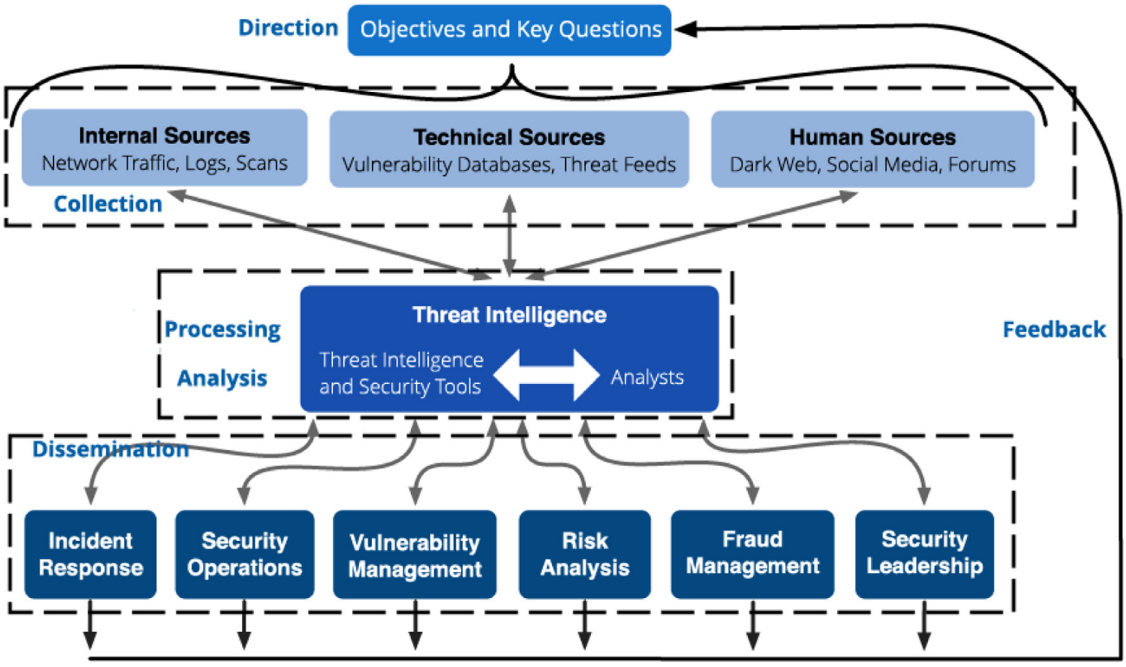


Fig. 1 – Threat intelligence lifecycle.

1. Introduction

1.1. *Vision and scope*

Techopedia[1] defines Cybercrime as "*[...] a crime in which a computer is the object of the crime (hacking, phishing, spamming) or is used as a tool to commit an offense (child pornography, hate crimes)*". Cybercriminals may use computer technology to access information that can be personal, business information like trade secrets, or use the net for any other malicious purposes. "Hackers usually perform these types of illegal activities". The damage that any single cybercriminal activity can bring about is massive. In 2017, the *WannaCry* ransomware was used to attack the National Health Service in May of that year, and Petya/NotPetya ransomware infecting global companies with a total waste of resources that cannot, to date, be estimated. The years 2018 and 2019 suffered no better fate, indeed (Solano and Peinado, 2021). The overly high costs connected to and lack of knowledge over these cyberattacks fundamentally motivates a systematic synthesis of the problem and solutions around the phenomenon. We operate such a systematic synthesis intending to identify gaps and shortcomings in the literature if any.

More specifically, we aim at investigating state of the art in *threat intelligence* systematically, that is, the discipline whose intent is that of providing organized, analyzed, and refined information about potential or current attacks that threaten an organization, including governments, non-governmental organizations, and more (Tounsi and Rais, 2018). There are six phases in order to make up the intelligence lifecycle (see Fig. 1) in such threat intelligence endeavors, namely: direction, collection, processing, analysis, dissemination, and feedback (Pokorny, 2020).

Direction is the phase where goals are set for the threat intelligence: *i* potential impacts of interrupting process (i.e., closing drugs markets or stop terrorist blogs), *ii* priorities about what to protect (i.e., stopping the arms trafficking in order to protect civilians life), *iii* The information assets and business processes that need to be protected. In this respect, we aim to identify the relevant dimensions of the problem under investigation in state of the art, thus allowing practitioners to identify what is available and researchers to identify what is missing and may need more research.

Collection is the process of information gathering: *i)* open-source scanning news, blogs, markets (i.e., retrieve user behaviour statistics from dark markets or apply thematic coding analysis to terrorist's blogs), *ii)* crawling and scraping forums, website, and any other relevant source, *iii)* infiltrating closed sources such as dark web forums. In this respect, we aim to identify the relevant techniques and policies to conduct such collection to simplify and support cyber-crimefighting practitioners' activity, e.g., Law-enforcement agencies.

Finally, concerning, **Processing**—the phase where all the collected data are formatted, filtered for false and redundant information, and made usable by the organization (i.e. extraction of IP addresses and creation of a CSV reporting file)—as well as **Analysis**—where the processed information is converted into intelligence by a human process, so that can inform decisions. Based on circumstances the decisions can involve the possibility to investigate a potential threat further, required action to stop an attack, how to improve security, eventually investments to take—and **Dissemination**—which involves getting the finished intelligence output to the places it needs to go. Threat information types include indicators, security alerts, threat intelligence reports, and tool configu-

[1] https://www.techopedia.com/definition/2387/cybercrime

ration information for using tools to automate all the phases of threat intelligence. Different intelligence reports are generated to meet the management and higher-level executives' requirements at strategic, operational, tactical, and technical levels.—Finally, we aim at collecting **Feedback** in order to understand intelligence priorities, focuses on what data is needed to collect, how to process this data in order to have some useful and usable information, and how to analyze data (Pokorny, 2020).

Given the extent of the phenomenon and the literature around it, we aim at accounting for both grey and white literature on the matter with a systematic *multi-vocal* literature review (Garousi et al., 2013; Kitchenham et al., 2008).

Our results provide a clear overview of the topics, approaches, indicators, risks, fallacies, and pitfalls around the phenomenon of threat intelligence.

1.2. *Approach and major contributions*

Overall, our work focuses on addressing five research questions: (i) Which online depth levels are assessed and to what extent?; (ii) Which degrees of anonymity exist for web-crawling?; (iii) Which policies exist to vary the degrees of anonymity?; (iv) Which website features are most indicative of cyber threats?; (v) Which risk assessment techniques exist? These five research questions come from focus groups and case studies with the Law Enforcement Agency (LEA), where we analyzed how the agencies work in cybercrime fighting and what tools they use to be effective but still working anonymously to protect the personal identity. Altogether, the research questions aim to shed light on the approaches and techniques that could be combined into a risk-assessment campaign enacted by law-enforcement agencies (LEAs) or policy-makers over cybercrime perpetrated in novel online sources. Our theoretical assumption is that offering this background is vital to enact an educated design of proper risk-assessment technology stemming from previous work in the field.

The significant contributions of this work are threefold: (a) a dual taxonomy for cyber-crime threat intelligence in the (a.1) surface web and (a.2) deep- and dark-web; (b) a rigorously mined set of topics that can be used as quantitative indicators for further risk assessment and confirmation; (c) a systematic overlap analysis between the just-mentioned contributions (a) and (b) to elaborated gaps in state of the art and opportunities for further research.

To recap the aim of this study, we offer below a summary of our pursued results.

Summary. Our survey study offers an overview of cyber threat intelligence, providing a taxonomy of the current criminal activities and complementary activities to detect, avoid, and assess them; more specifically, we provide an overview of indicators and risks parameters in order to aid Law-Enforcement Agencies in their cybercrime fighting activities. This is the first systematic literature review with this perspective over cyber threat intelligence to the best of our knowledge.

The rest of this paper is organized as follows. First, Section 2 outlines the background as well as related work. Further on, Section 3 elaborates on the research design behind this study. Subsequently, Section 4 outlines the results while Section 5 discusses them in context. Finally, Section 6 first introduces a Research Roadmap then concludes the paper.

2. Background and related work

To the best of our knowledge, this is the first systematic literature review providing a taxonomy about the different types and dimensions of cybercrime and threat intelligence solutions. However, in the following paragraph, we discuss papers that provide a partial overview of threat intelligence rather than existing literature that solve the problem of cybercrime risks or guidance notes to assist in addressing the problem posed by cybercrime. In the available literature, no surveys are trying to create a general overview of the cybersecurity risks and the proposed solutions to contain the risks. Our systematic literature review analyses the state of the art of upcoming cybersecurity risks and the proposed countermeasures today available. Due to the novelty of the cybersecurity threats and the lack of technologies available to fight the cyberattacks, we will also examine sources from the web like blogs and news to have a broader perspective on the new cybercrime trends.

2.1. *Related surveys*

In a deeply connected world, like the one we are facing nowadays, hackers continuously try to find new targets and develop new tools to break through cyberdefenses. Moreover, the lack of privacy and security of the new upcoming technologies and the users' lack of awareness pose a real threat to our personal life. In the following, we present some works that face cybersecurity and discuss the countermeasures today available.

Tounsi and Rais (2018) provides an overview of the open-source/free threat intelligence tools and compare their features with those from AlliaCERT TI.[2] Their analysis found that the fast sharing of threat intelligence (i.e., Fernández Vázquez et al., 2012; Jasper, 2017), as encouraged by any organization in order to cooperate, is not enough to avoid targeted attacks. Moreover, trust is essential for companies that are sharing personal information. Another problem is how much data is necessary to share to prevent attacks and cooperate and in which format to avoid losing information. In order to understand which standard is better Tounsi et al. propose their analysis. Lastly, the work compares the best threat intelligence tools that divide them into tools that privilege standardization and automatic analytics, and others that focus on high-speed requirements.

Furthermore, if Tounsi et al. focus on what is the best way to keep the trust among organizations and at the same time share information about cyber threats, in Toch et al. (2018) the authors place the accent on the type of data required from those cybersecurity systems that are supposed to protect our privacy from prying eyes. The taxonomy suggested in the article lists the risks of the different types of cybersecurity technologies related, which are related to a specific cyberattack.

[2] Managed Security Services Division, AlliaCERT Team, Alliacom, France.

The taxonomy shows that almost all cyber-security technological categories require some access to personally sensitive information. This result can offer guidance not only in choosing one technique over another but, more importantly, in designing more privacy-aware cyber-security technologies with little or no compromise concerning their effectiveness in protecting from cyber-attacks.

The studies from above tried to analyze systems and good practices to mitigate cyber threats. In Chang et al. (2013), we have a study regarding the state-of-the-art web-based malware attacks and how to defend against them. The paper starts with a study about the attack model and the vulnerabilities that enable these attacks, analyzes the malware problem's current state, and investigates the defense mechanisms. As a result, the paper gives three categories of approaches in order to analyze, identify, and defend against the web-based malware problem: (1) building honeypots with virtual machines; (2) using code analysis and testing techniques to identify the vulnerabilities of Web applications; and (3) constructing reputation-based blacklists. Each category with advantages and disadvantages, how these approaches complement each other, and how they can work together.

An altogether different approach from the previous ones is presented in Xu et al. (2013) where the authors analyze network-layer traffic and application-layer websites contents simultaneously in order to detect the malicious web applications at run-time. The currently available approaches to detect malicious websites can be classified into two categories: *static approaches* and *dynamic approaches*. The first approach analyzes URLs and contents; the latter uses clients honeypots to analyze run-time behaviours. The experiments with this approach showed that cross-layer detection could achieve the same detection effectiveness as the dynamic approach. However, it resulted in being much faster than the dynamic one.

In order to understand the rising concern around the cybersecurity problem, another important reference is also the *Guidance Note*[3] of the United Nations Office on Drugs and Crime (UNODC) that is a global leader in the fight against illicit drugs and international crime. The guidance note aims at giving a comprehensive overview of the most common cybersecurity threats today available. To outline how UNODC can deliver technical assistance to address cybercrime problems at both regional and national levels, cybercrime activities like online radicalization or the illicit sales of pharmaceutical solutions are presented and explained in detail. If we see a considerable growth of interest around cybersecurity threats from the scientific literature side, we have many blogs and webpages warning about the new upcoming cybersecurity threats on the web side.

Furthermore, we refer to reports of one of the major companies working in cybersecurity: Kaspersky.[4] On the Kaspersky Threats blog page,[5] where the company offers an updated list of the new upcoming cyber threats. More specifically, on the top five worst cybersecurity attacks, we have WannaCry and NotPetya/ExPetr, two famous ransomware encryptors used to encrypt the victim user's data. The worm Stuxnet, the spyware DarkHotel, Mirai, a botnet used to flood the DNS service provider. The Kaspersky company gives guidelines (Kaspersky Lab daily, 2018) on how to address incident response to contain a cybersecurity attack. Kaspersky listed some key-points necessary for a company to avoid and contain attacks: *(i)* the speed, rapid remediation is a key to limiting the costs, *(ii)* proactive protection, *(iii)* presence Of internal specialists. However, to have a worldwide overview of real-time cyber-attacks, Kaspersky provided the Cyber threat Real-Time Map available here https://cybermap.kaspersky.com/ where it is possible to see the current cyber-attacks around the globe.

Altogether, however, although plenty of white/grey literature exists on the topic, a holistic view over what software, indicators, methods, tools, and approaches to cyber-crime fighting that practitioners and law-enforcers can use is still nowhere to be seen. We offer an initial attempt at such a review in the next pages to benefit practitioners and academicians alike. Another tech player working in cybersecurity is Norton.[6] In (Symantec Employee, 2018), a Symantec employee gives a picture of cybersecurity threats and their impact on the American population. Mobile malware and third-party app stores seem to be a new concern. If, until 2017, spyware, ransomware, and viruses were focusing mainly on laptop and desktop PC, after 2017, has been recognized an evolution of malware attacks for mobile and an increment of 54%. From the Symantec report, in 2023, cybercriminals will be able to steal something like 33 billion records that might include names, addresses, credit card information, or Social Security numbers. The impact of this identity theft will impact 60 million Americans, and the average costs have been estimated at $3.86 million (U.S. dollars) for the companies worldwide and $7.91 million (U.S. dollars) for the U.S. company.

3. Research materials and methods

This systematic literature review seeks to address the research problem of providing a clear and detailed overview of the methods and indicators used for cybercrime threat intelligence. Because much work has been conducted and disseminated in non-scientific venues and by non - governmental organizations, we opt for a systematic *multivocal* literature review (Garousi et al., 2017), meaning that both grey and white literature are considered as equal sources of valuable data. In the rest of the section, we flesh out the research questions and methods we employed to attain our results.

3.1. *Research problem, questions and motivations*

The methodology used to attain our results is described in the following paragraphs and is tailored from our previous work (Soldani et al., 2018).

More specifically, we seek to address the following master research question (MRQ):

MRQ. *what guidelines, methods, and principles exist to establish the cyber threat level of online sources?*

[3] https://bit.ly/2BIy0tP
[4] https://www.kaspersky.com
[5] https://bit.ly/2AijYiF
[6] https://us.norton.com/

The question mentioned above can be rephrased into the sub research questions (SRQ) reported here below:

(SRQ 1.) what online depth levels are assessed and to what extent?
(SRQ 2.) what degrees of anonymity exist for web-crawling?
(SRQ 3.) what policies exist to vary the degrees of anonymity?
(SRQ 4.) what website features are most indicative of cyber threats?
(SRQ 5.) what risk assessment techniques exist?

As above mentioned, the five research questions were elicited in five focus groups held with 2–3 practitioners from six LEAs equally distributed across Europe, encompassing the citizenships of Brecht (BE), Sofia (BG), Eindhoven (NL), Belgrade (SLO), Gdansk (PO), London (UK). More specifically, in the context of EU H2020 project[7] we investigated the tools and techniques used to fight cybercrime among the LEAs mentioned above as well as 5 SMEs (Small and medium-sized enterprises) and 2 NPOs (a non-profit organization) from eleven different countries. All the involved LEAs and practitioners are actively engaged in software technology and engineering management approaches to cyber crime-fighting and cyber-threat intelligence. The five research questions resulted as the most prioritized items out of a card-sorting game (Lewis and Hepburn, 2010) based on the work carried out by law-enforcement agencies, and in order to cover all the different types of investigation they are conducting daily.

The SLR contained in this manuscript takes the RQs as a basis for investigation and covers all the aspects and technologies of cybercrime fighting expressed in said RQs, namely, (1) the tools used to perpetrate an attack, (2) what are the available countermeasures, (3) how to crawl the web, (4) till the policies to protect the identity of the police officers that are investigating.

More in detail, in terms of SRQ1, we aim at figuring out which analysis techniques exist that cover which level of depth (i.e., what analysis exists for the Surface web and with what grade of detail). Furthermore, in the scope of SRQ2 and SRQ3, we aim at understanding the techniques and approaches that would allow a law-enforcer to crawl online sources anonymously and to what extent this phenomenon is understood and addressed in the literature. Beyond that, with SRQ4 and SRQ5, we aim at figuring out which detection and analysis techniques exist and how they can be used, that is, upon which data features (Zave, 2003). This study's significant intrinsic difficulty is our necessary reliance over what is called grey literature (Garousi et al., 2016), intended as materials and research produced by organizations outside of traditional commercial or academic publishing and distribution channels. Common grey literature publication types include reports (annual, research, technical, project, etc.), working papers, government documents, white papers, and evaluations. On the one hand, grey literature usage is risky since there is often little or no scientific factual representation of data or analyses presented in grey literature itself (Farace and Schöpfel, 2010). On the other hand, a growing interest around using grey literature for computing practitioners' benefit as well as combining it to determine state of the art and practice around a topic is gaining considerable interest in many fields (Farace and Schöpfel, 2010; Stempfhuber et al., 2008), including software-related fields (Garousi et al., 2016). For the scope of this paper, and to maximize its validity, it follows a systematic approach based on the guidelines provided by (Petersen et al., 2008) for conducting systematic literature reviews in software engineering. Is hereby outlined such a systematic approach, starting from problem definition and describing the triangulation as well as other inter-rater reliability assessment trials, we ran to enforce the validity of our findings. More specifically, the following search query was derived directly by isolating the keywords in our RQs and using the approach identified by Farace and Schöpfel (2010):

$$(\texttt{cyber*} \vee \texttt{online*}) \wedge (\texttt{threat*} \vee \texttt{attack*} \vee \texttt{activity*} \vee \texttt{crime*}) \wedge (\texttt{Surface*} \vee \texttt{D*})$$

In the above, the "*" symbol is the star wildcard that matches lexically-related terms (e.g., plurals, verb conjugations). We narrowed results obtained from the search string to industrial, government, and non-governmental studies (e.g., blog posts, white papers, industry-oriented magazines) published from the beginning of the internet (early 90's) until the mid of 2018. The search engines we decided to use are Google (primary) and Bing. Our search resulted in a high number of irrelevant studies (60%). This is because the search engines look for the above-indicated search strings over the whole pages they index. Moreover, the research has been further refined with a secondary search and manual screening, based on the inclusion/exclusion criteria and control factors discussed in the following section.

At the same time, to cover for white literature appropriately, we run the query as mentioned earlier in typical and most common computing literature libraries, namely: (1) ACM Digital Library; (2) IEEEXplore; (3) Wiley Interscience; (3) Elsevier Scopus; and (4) Bibsonomy.

3.2. *Sample selection and control factors*

The criteria (inclusion and exclusion) adopted in order to build our sample selection are outlined in Table 1

The inclusion criteria ($i_1 - i_4$) were designed to focus on that side of the grey literature that identifies the targets of our study. On the other side, the exclusion criteria permit disqualifying studies that do not offer the necessary design/implementation details (e_1), that refer to unquantifiable evidence (e_2 and e_3), and that do not examine the limitations and practical impact for the proposed solutions or the presented issues (e_4 and e_5). To select a study, it needs to satisfy *all* the inclusion criteria, while it is excluded when it satisfies at least one of the exclusion criteria.

In addition to the inclusion/exclusion criteria in Table 1, to ensure the quality of the selected grey literature, we selected only those industrial studies that are satisfying the following control factors:

1. **Practical experience.** A study is to be selected only if it is written by practitioners with 5+ experience in the topic in object, or if it refers to established threat intelligence solutions with 2+ years of operation.

[7] https://www.anita-project.eu/

Table 1 – Inclusion and exclusion criteria for sample selection.

Case	Criteria
Inclusion	i_1) The study discusses cybercrime or an application of analysis to the topic.
	i_2) The study discusses the ramifications and challenges around the topics close to our RQs.
	i_3) The study address know-how, opinion, or practices on the topics in our RQ by directly-experienced practitioners.
	i_4) The study reports a case-study of cybercrime threat intelligence incidents or approaches.
Excl.	e_1) The study does not offer sufficient details on the design or implementation of practices, methods, tools, or cybercrime threat intelligence indicators.
	e_2) The study is not referred to practical cases or does not report any factual evidence.
	e_3) The discussed topics are not justified/quantified by the study.
	e_4) The study does not discuss the scope and limitations of proposed solutions, frameworks, patterns, tools.
	e_5)The study does not offer evidence of a practitioner perspective.

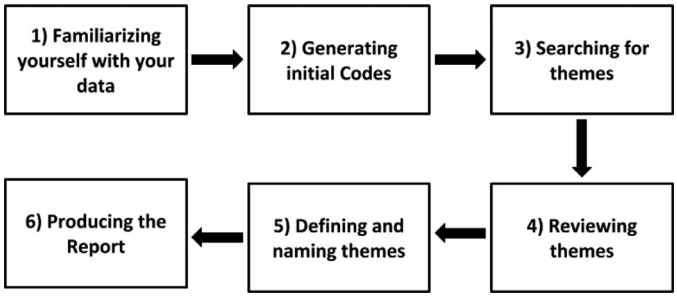


Fig. 2 – 6-phase thematic coding analysis.

2. **Industrial case-study.** A study is to be selected only if it refers to at least one industrial case-study where a quantifiable number of threat intelligence tools are operated.
3. **Heterogeneity.** The selected studies reflect at least 5 top industrial domains and markets where threat intelligence tools were successfully applied.
4. **Implementation quantity.** The selected studies refer to/show implementation details for the benefits and pitfall they discuss so that other researchers and practitioners can use them in action.

At the end of our screening, 374 studies were selected based on the inclusion/exclusion criteria. For the screening process we used not only the inclusion and exclusion criteria but also the application of our quality control factors (1. article is written by practitioners with 5+ experience, 2. article refers to at least 1 industrial case-study, 3. selected studies reflect at least 5 top industrial domains). The complete list of selected studies is provided online[8].

3.3. *Data analysis*

To attain the findings, a mixed-method analysis approach was adopted (Johnson and Onwuegbuzie, 2004). First has been applied Thematic Coding in order to generate themes from the involved data. Here below, Fig. 2 gives a graphical representation of all the phases of the analysis process.

Below the explanation of all the phase as reported in Clarke and Braun (2013)

- **Familiarization**. We started reading the papers in order to become familiar with the collected data.
- **Generating the initial codes**. After becoming familiar with the data, we started the coding stage. In this phase, it is important to isolate those phrases, sentences, and paragraphs related to our topic. We analyzed all the papers and extracted phrases, sentences, and paragraphs to create clusters of themes.
- **Create the initial themes**. In this phase, we revised the codes clustering them together if there are similar meanings or found relationships among them. This stage helped us in identifying patterns among codes.
- **Review the initial themes**. In this phase, we had to ensure that our themes are useful and accurate representations of the data. Hence, we return to the data set and compare our themes against it.
- **Name and define the themes**. In this phase, it is essential to create proper labels to develop a succinct and easily understandable name for the different themes involved.
- **Write the final report**. After the definition of the themes and naming them, we started writing our final report.

Finally, to address SRQ4, we operated a machine-assisted topic modelling and analysis exercise supported with thematic coding. We chose the Latent Dirichlet Allocation (LDA) approach since it is the most popular and generally the most compelling topic modelling technique. With LDA, we can extract human-interpretable topics from our data set of papers where each topic is characterized by those words that are most strongly associated with. We used Latent Dirichlet Allocation (LDA) to provide emerging themes in our textual data, subsequently labeling the emerging themes that are visible and ob-

[8] https://tinyurl.com/ANITAstudysourcesMSLR

Fig. 3 – Types of organizations involved in the grey-literature, a majority of public organizations are involved.

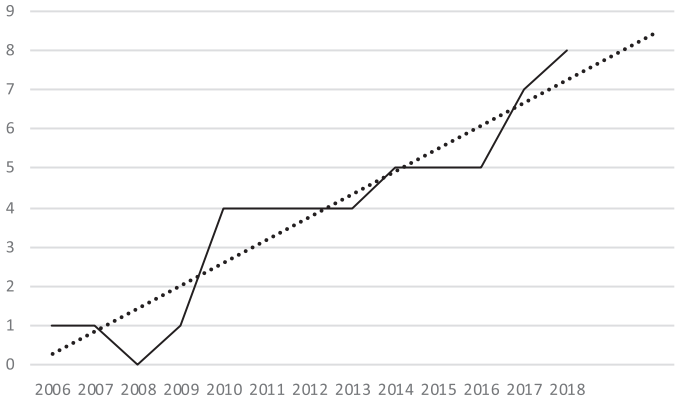


Fig. 4 – Increase of interest over the topic; a linear increase is reported.

servable characteristics of potential online sources for criminal activity (e.g., darknet websites).

For the afore-mentioned topic modelling exercise, we selected log-likelihood as our measure of clustering quality, following typical approaches from state of the art (Agrawal et al., 2016). In our case, however, the number of clusters started from typically used numbers adopted in state of the art ($k = 10$ clusters), but the number was increased until at least one of the newly-emerging clusters contained less than half of the mean population of factors in the previous round. This approach aimed to allow the extraction of cybercrime activities and meaningful indicators, i.e., they reflected semantic commonalities among factors. Besides, we used the genetic algorithm Differential Evolution to tune LDA hyperparameters alpha and beta, as suggested by (Agrawal et al., 2016). To conduct all the above pre-processing and analyses, we exploited the NetCulator bibliometric analytics tool,[9] which supports LDA and several similar natural-language analyses and clustering techniques and tools of our design, featuring Python and the python LDA package.

[9] https://www.netculator.com/

4. Results

This section outlines study results from descriptive statistics and later elaborates on the taxonomy of concepts found for surface-web cybercrime threat intelligence and later elaborates on taxonomy of concepts found for deep- and dark-web cybercrime threat intelligence. For the deep- and dark-web taxonomy, all concepts reported in the surface-web taxonomy were also found but were omitted for the sake of completeness and to highlight more specific results.

4.1. *Data set descriptive statistics*

The descriptive statistics below are brief descriptive coefficients that summarize our data set. The descriptive statistics help describe and understand our data set's features by giving short summaries about the involved data's sample and measures. It hence contributes to the final key findings giving an overview of the data and the data set itself we used to carry out our research.

The primary sources we reported offer a diverse statistical distribution over the last 20+ years. Figures from 3–5 outline statistical descriptors for the elicited grey literature while Figs. 6–8 offer a similar insight into white literature. More specifically, Fig. 3 outlines the types of organizations

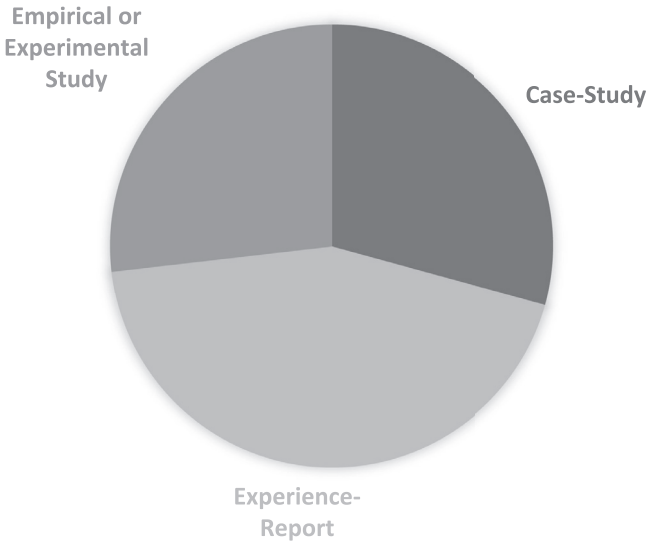


Fig. 5 – Types of evaluation involved in the grey-literature; experience reports are the striking majority.

that conducted the research reported in our primary studies, ranging from private corporations (e.g., Kaspersky labs) to public institutions (e.g., non-governmental organizations and boards), who cover for the majority of our sample. Further on, Fig. 4 provides a timeline reflecting a linear increase in interest over the phenomenon between the oldest (2006) and newest (2019) article we analyzed. From Fig. 4, we can infer how, over time, the problem of cybercrime threat intelligence becomes more relevant for both private corporations, public institutions, and academics. While Fig. 5 provides a deeper insight into the types of evaluations conducted in the grey-literature in question, with a striking majority of experience reports being used as a basis for argument. An experience report is a paper written by a person (persons) that systematically report the experience reported through direct experimentation.

On the white-literature front, Fig. 6 offers an overview of the types of studies reported in the literature, with a majority of case-studies being targeted for further research. A case study is a research strategy and an empirical inquiry investigating a phenomenon within its real-life context.

Beyond the types of studies, Figs. 7 and 8 offer an overview of the topic interest — which reflects some mixed trends — and the typical venues, with a striking preference for conferences — which are typically more divulgative in nature.

Overall, the statistics offer a not-so-comforting picture. The field seems in an emerging phase, with mixed-feelings or forming interest, typically disseminated in conferences but discussed over case-studies (in white) and/or from experience reports (in grey literature). More specifically, the data reveals a linearly increasing trend, but the Chart in Fig. 4 remarks about the relative lack of large-scale experimentation.

Finally, Fig. 9 offers a quantitative overview of the core concepts discovered as part of our analysis (Definition of codes is provided). The figure highlights that most of the literature we analyzed focuses on discussing specific detection *methods* for *criminal activity types*, as opposed to providing holistic methods for the discovery of cybercrime. Moreover, from a quantitative perspective, we highlight that *website appearance* and their degree of (software) *security* are major indicators for risk assessment. The next sections offer more details on the results of our study.

4.2. *Cybercrime threat intelligence: a surface-web taxonomy*

Fig. 10 (and later, Fig. 11) outlines the result of our thematic coding as applied to literature discussing or targeting analyses on the *surface* web only. In synthesis, both taxonomies address the current literature in cyber threat engineering and management concerning this manuscript's research questions. The text below outlines and illustrates the taxonomies and connected research results.

The results are articulated using a simple UML-like model structured using the core-concepts (inner-most, white boxes on Fig. 10) emerging from our thematic coding, namely: (a) *assessment methods* — these are the methods, techniques or tools discussed in the state of the art to address cybercrime threat intelligence; (b) *countermeasures* — these are the methods and measures that can limit the damage connected to cybercrime, as discussed in literature; (c) *anonymous crawling policy* — these

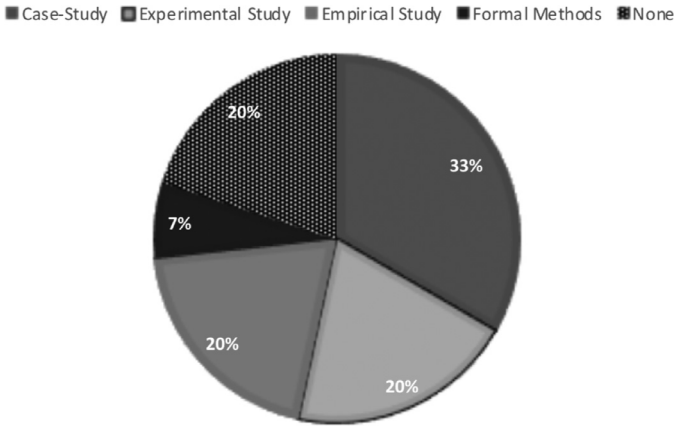


Fig. 6 – Types of studies conducted in white-literature; case-studies are targeted the most.

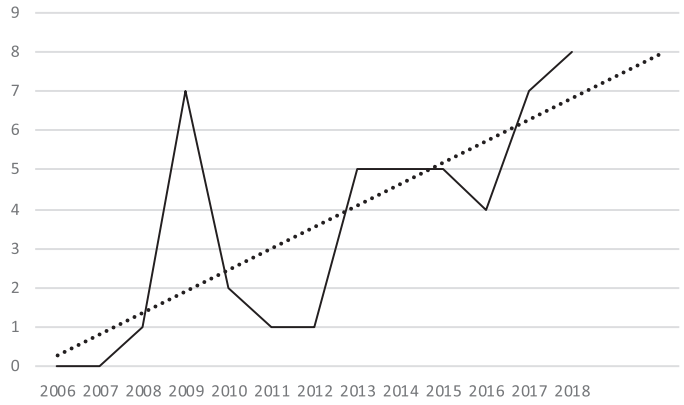


Fig. 7 – A linear trend is present in white-literature as well; however, mixed but rising interest is reported over the years.

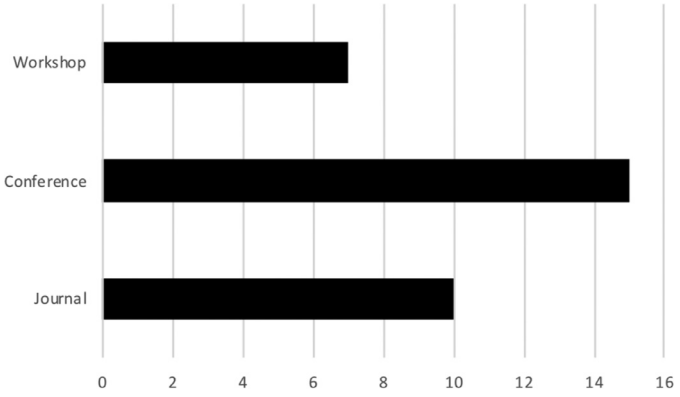


Fig. 8 – Venues selected for publication; the strong preference for conferences or workshops as opposed to journals reflect an emerging discipline.

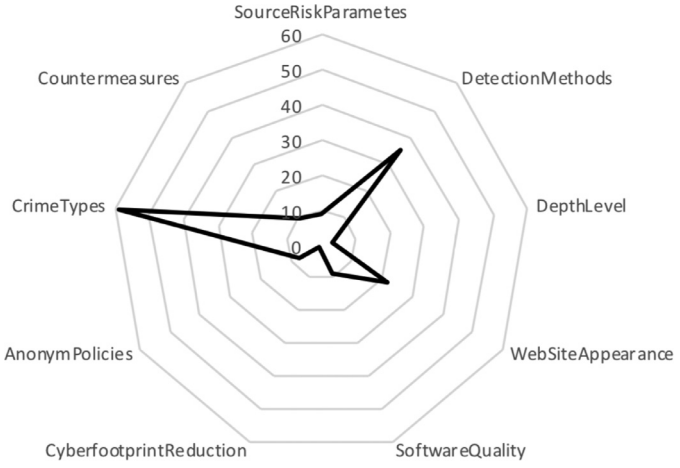


Fig. 9 – Count of occurrences for core-concepts across our data set, normalized on a percentile scale.

are the techniques and policies that can limit the detection risk of conducting cybercrime threat intelligence in the open; (d) *risk-level parameters* — these are indicators for increased risk of specific cybercrimes; (e) *website appearance parameters* — these are "hints" that previous research identifies as a certain factor indicating that a web source is hosting a specific criminal activity; (f) *software-quality parameter* — these are software-related quality metrics (e.g., increased throughput or reduced responsiveness) that indicate or are connected to a specific criminal activity being perpetrated; (g) *criminal activity type* — these are the actual criminal activities being carried out.

The outer-most, grey-colored boxes on Fig. 10 outline what we reported from literature, with a frequency cut-off of 3 re-

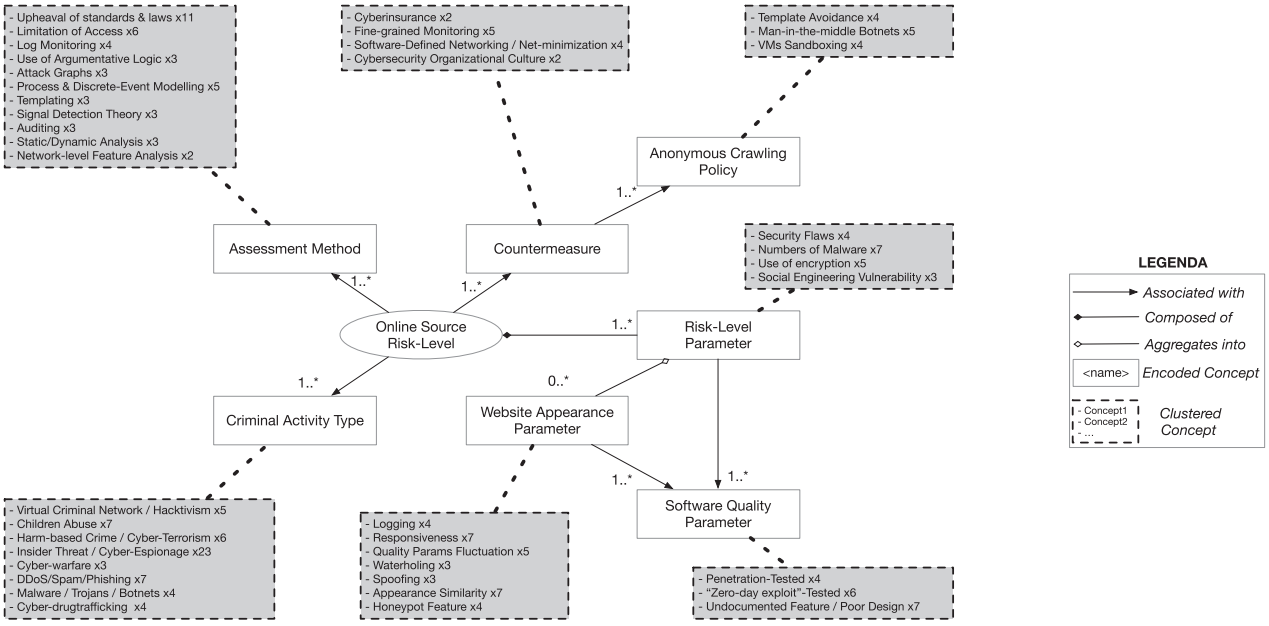


Fig. 10 – A taxonomy of cybercrime threat intelligence for the surface web.

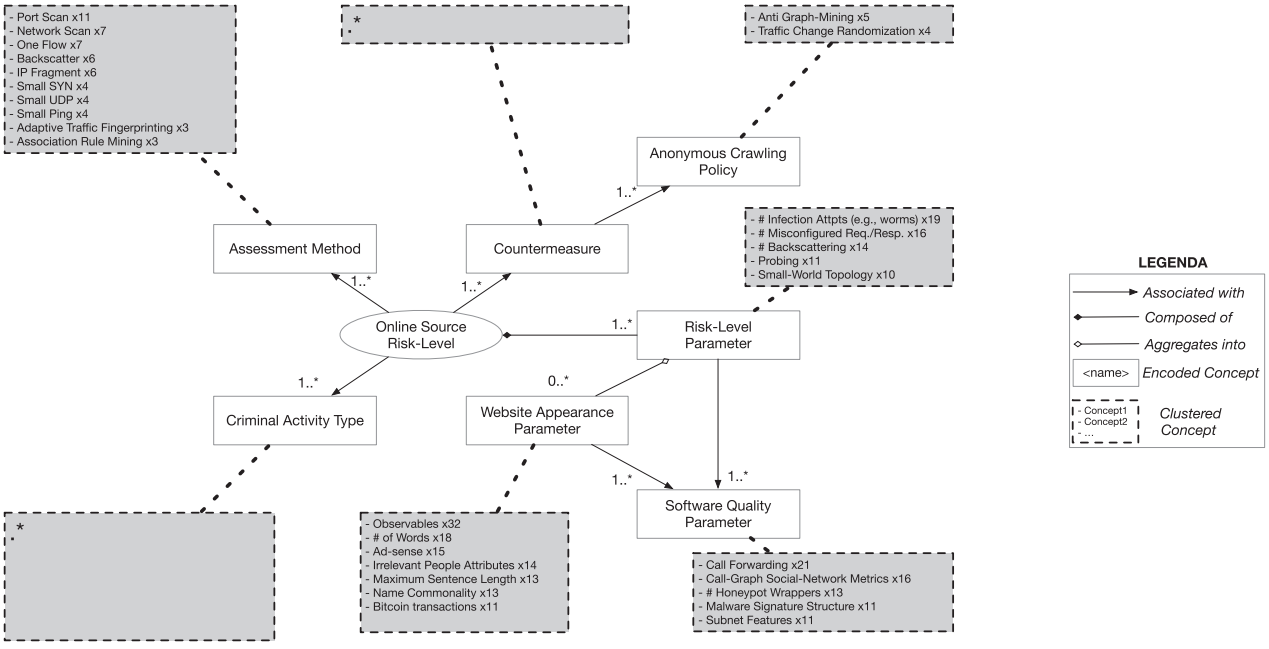


Fig. 11 – A taxonomy of cybercrime threat intelligence for the Deep-, Dark Web.

currences over three primary studies from *both* grey and white literature, meaning that concepts, techniques, tools, and methods discussed less than three times and published or discussed before 2018 were not reported for the sake of space.

In the following, we flesh-out the results from Fig. 10 in the same order as the core-concepts were outlined in the text above; resulting concepts appear in *italics* in the descriptive sections. It should be noted that, from this point forward, no distinction is made between grey or white literature to avoid any bias in the exposition of the results.

4.2.1. *Assessment methods (METH)*

From a policy perspective, literature remarks that the use of *standards and laws* is the single most-used risk assessment method against cybercrime activity; several articles in both grey and white literature remark that the Gramm-Leach-Bliley Act (GLBA) (Chen et al., 2004) or the Fair Credit Reporting Act (FCRA) (Hoofnagle, 2013) offer the technical and legal basis to establish the perpetration of online financial crimes of multiple types. In over 30% of our sample, similar legislation (including GDPR in more modern instances) are suggested as tools in their own right to be used against cybercrime of a

more shallow and evident nature in the surface web. Furthermore, several experience reports and case-studies elaborate on the use of *limitation of access* or access-control blacklists as a method to establish and limit the involvement with cybercrime. More specifically, tools and approaches such as SquidGuard [10] offer a basis to share and adopt lists of sites hosting criminal activities to be avoided.

From a more technical perspective, *log monitoring* is highlighted as the most obvious cybercrime risk detection and avoidance method. Mataracioglu et al. (2015) report on a cybercrime and cybersecurity framework which harnesses log monitoring to detect and avoid social engineering tactics often employed as part of cybercrime. A similar argument is made for the use of log monitoring in several articles from the proceedings of the federated conference on Data Privacy Management, Autonomous Spontaneous Security, and Security Assurance (García-Alfaro et al., 2015). In these venues, log monitoring is combined with *attack graphs*, a formalism built on top of log monitoring techniques that can elicit social engineering attacks by dissecting the connected social engineering threats and vulnerabilities (Beckers et al., 2014). Similarly to attack graphs, log monitoring and similar runtime threat detection and avoidance activities combine *process modelling/mining* and *argumentative logic*. Bouyahia et al. (2014) introduce a metrics-based technique to assist the detection and avoidance of security threats using reasoning systems that incrementally figure out ongoing attacks — while ontology-based approaches are highlighted in the paper, the authors also remark on the potential to combine a more data-driven machine-intelligence approach.

From a process mining and modelling perspective, the techniques of *discrete event modelling* dating back to '97 and to Harel and Gery seminal work on object statecharts (Harel and Gery, 1997), to signals-detection theory (Green and Swets, 1989) and signals intelligence (Ma et al., 2018) applied to *static/dynamic networks traffic analysis* and ending up with a recent work focusing on terrorist attacks by Gabriel et al. (2017).

Overall, the state of the art results as *very* domain-specific (e.g., terrorist attacks Gabriel et al., 2017, insider threats Blackwell, 2009) mostly based on *templating* of crimes — that is, offering a standardized format for the perpetrated crime and matching that format onto available data — and with little generalizable approaches.

On the other hand, the last two approaches we reported as recurrent, namely, *auditing* and *network-level feature analysis*, offer theoretical bases for generalisability. More specifically, cybercrime auditing entails providing for strategic checking of organizational and technical infrastructures by randomly selecting a cybercrime type, instrumenting the type, and purposefully targeting the organizational and technical infrastructures with it to evaluate the target infrastructures' vulnerability to it.[11] Concerning the auditing technique highlighted above, Chang et al. (2013) offer a more thorough overview of malware-based crimes which is offered as a basis for targeted auditing.

Finally, concerning network traffic analysis, several approaches reported in literature offer feature-based (social) network analysis (O'Riordan et al., 2016), as well as feature engineering and analysis techniques aimed at establishing precursors of social engineering, most notably from our data set the works by Vidal and Choo (2017) or Gharibi (2012a).

4.2.2. *Countermeasures (CMEASURE)*

As previously specified, with the term *countermeasure*, we identify the ability to foresee and enact preemptive or corrective action against a specific cybercriminal activity. Most of the grey literature highlights the need to conduct a business-level impact assessment and incident management. The report of the Australian Government (Ring et al., 2017) remarks that businesses need to be arranged, quoting from the original document, specific *"actions taken as soon as an attack or breach has occurred to determine the (1) depth of its effect on the business, (2) your ability to recover, and (3) affect the likelihood of future breaches"*. Several proactive actions have been introduced. Baer et al. discuss several approaches to Cyber insurance (Baer and Parkinson, 2007), and similarly, earlier works by Meland et al. (2015) establish the ways in which cyber insurance actions can be planned as part of corporate governance and towards the reduction of cyber threats risks.

From a more analytical perspective, several technical countermeasures were proposed, mostly along the lines of fine-grained monitoring of IT assets and business processing. More specifically, Ma et al. (2012), as early as 2012, offer a lightweight framework for monitoring public clouds, which are outlined as a potential solution for mitigating cyber threats, as long as an appropriate incident response organizational structure and culture (Chang and Lin, 2007; Tang et al., 2016) is also in place whereupon a threat does manifest. Later works offer prototypical solutions where cloud and IT infrastructures monitoring is combined with real-time applications security (Coppolino et al., 2014). Still, on a technical perspective, acting as a countermeasure for cybercrime is the use of software-defined networks (SDNs) as well as virtual-networks functions (VNFs), that is, harnessing with programmable / controllable software the responsibility of handling specific network functions that run on one or more virtual machines. In this specific domain, the survey by Scott-Hayward et al. (2013) offers an overview of the practices in SDNs, which can be used to attain software-controlled granular cybersecurity and safety.

4.2.3. *Anonymous crawling policies (ACP)*

In terms of maintaining anonymity while performing cybercrime detection or avoidance tasks across an organizational structure, much research has devoted to using and refining Bots and botnets dedicated to detecting social engineering attacks or performing anonymous analysis. The works by Lauinger et al. (2010a) and subsequent trials by the US Chamber of commerce contained in their whitepaper[12] remark that *"an acceptable-use policy for the use of information resources and IT systems [needs] for example, confidential or sensitive business information not to be posted by employees on social networking sites such as Facebook or MySpace [...]"*; the aforementioned actions were

[10] http://squidguard.mesd.k12.or.us/

[11] http://m.isaca.org/knowledge-center/research/researchdeliverables/pages/cybercrime-audit-assurance-program.aspx

[12] https://www.uschamber.com/CybersecurityEssentials

experimented upon with the usage of policy-driven bots to perform counterinsurgency of amended actions. Likewise the survey by Chang et al. (2013) offers an overview of several approaches along the lines defined above, wherefore web-based malware is detected, risk-assessed, avoided using on-purpose, policy-driven botnets.

Finally, in terms of anonymity during detection phases for cybercriminal activity, the use of Virtual-Machine sandboxes is often referred to as the only viable mechanism (Chang et al., 2013). However, several recent works show the endurance of specific attacks or other masqueraded cybercriminal activity such as the S$A and similar shared-cache attacks (Apecechea et al., 2015) against a sandboxing approach.

4.2.4. *Risk-level parameters (SRLP)*

This section showcases the few parameters reported in the literature that are commonly known to increase cybercriminal activity risks being perpetrated in targeted online sources. An outstanding number of whitepapers and governmental reports highlight the presence and proliferation of several risk-related parameters. As noted in the US Chamber of commerce whitepaper about cybercrime,[13] "[actions need to be taken to] root out security flaws in computer programs and to counter cyberattacks by "bad" hackers, or cybercriminals". Moreover, the US Chamber indicated the presence and extent of security flaws (of which, the number of Malware is an established minimum, as noted by Rahul and Sujata, 2018 and several others Caballero, 2012) in the code of online sites as a probable factor of risk in establishing high-threat sources. Finally, the haphazard use (or lack thereof) of encryption across online source functions has been established to lead to cybercriminal activity, most notably in the roadmap defined by Kieseberg et al. (2015). More specifically, the lack of encryption is often connected to the use of specific social engineering activities being perpetrated in online sources, which themselves are functional to cybercrime (Gharibi, 2012b). On this latter front, that of social engineering vulnerabilities specifically designed to accommodate for cybercriminal activity, several authors such as Vidal and Choo (2017) remark on the necessity to conduct scenario-based situational crime prevention, e.g., using evolutionary computing and social predictive analytics — the work along these lines has mostly concentrated on elaborating more or less complete cyber forensics ontologies for the purpose of knowledge representation and reasoning about cybercriminal investigation in a scenario-based fashion (Park et al., 2009).

4.2.5. *Software quality parameters (SQUAL)*

The necessity to establish security as a software quality parameter to decide whether an online source bears risk of cybercriminal activity finds agreement in 90% of both grey and white literature alike. More specifically, the quality of software security is established around three axes: (1) whether the online source bear signatures and certificates of successful penetration-testing (Franklin, 2018); (2) whether the online source has been certified against morphisms (Gupta and Rani, 2018; Li et al., 2006) of known zero-day exploits (Bilge and Dumitras, 2013; Danforth, 2011); (3) finally, whether the online source bears undocumented software features and/or the indications of poor design (e.g., technical debt, etc.) (Nord et al., 2016).

4.2.6. *Website-appearance parameters (WSAP)*

In terms of website appearance, the literature we analysed identifies seven features as indicative pre-conditions to cybercriminal activity: (1) the lack of logging as well as software features for forward error correction, site responsiveness as well as other constructs that measure all graph-theoretic properties of the darknet (e.g., see Griffith et al., 2017); (2) variable responsiveness rates from the online source (3) a heavy fluctuation of the overall software quality parameters (e.g., language clarity, documentation, feature stability, etc.) for the online source, oftentimes detected thorugh anomaly detection or linear-time temporal logics, as seen in Almukaynizi et al. (2018); (4) the existence of waterholing features, defined by Trendmico[14] as areas of the site which are uncontrolled, uncontrollable, or never improved overtime by site maintainers (Khan, 2006); (5) the presence of spoofed information mismatches detectable through online fact- checking, an approach to this is presented in Nunes et al. (2018); (6) a high degree of appearance similarity with respect to other known online sources (Ghosh et al., 2017; Martine and Rugg, 2005); (7) finally, honeypot features most predominantly the length and target of the redirection chain upon any navigation request from the source, since almost 68% of our sources from white and grey literature studies observe that malicious landing sites almost always have unusually long redirection chains toward malware distribution sites (Chang et al., 2013).

4.2.7. *Criminal activity types (CTYPE)*

Lastly, the risk assessment of online sources can be supported by focusing on identifying the risk using combined measures of the likelihood for reported criminal activity types (Elstob, 1974). This section outlines and discusses all criminal activity types we reported in the literature. As previously remarked, we report in this section the crime types reported at least three times in at least three papers from both grey and white literature (i.e., at least six papers in total), later in Section 5 we discuss emerging crime activity types reported in more recent literature. Overall, the literature on cyber threat intelligence focuses around seven criminal activity types, namely: (1) Virtual Criminal Network / Hacktivism Groups — these reflect, on the one hand, crime networks dedicated to regular crime activity (e.g., drug trafficking) exploiting online means (Han et al., 2017) and, on the other hand, forms of cyber-activism (i.e., Hacktivism), where cyberattacks are ideologically motivated; (2) Children Abuse — these reflect sites exploiting minors for malicious intents and purposes, including and not limited to humans trafficking (Han et al., 2017); (3) Harm-based Crime / Cyber-Terrorism — these activities are usually ideologically motivated, as outlined by Gordon (Gordon and Ford, 2002)), and try to influence a state or an international organization exploiting system vulnerabilities (Veerasamy and Grobler, 2015); (4) Insider Threat / Cyber-Espionage — these activities focus on the exploitation of or-

[13] https://www.uschamber.com/CybersecurityEssentials

[14] https://www.trendmicro.com/vinfo/in/threat-encyclopedia/web-attack/137/watering-hole-101

ganizational insiders (Rocha, 2015) for information trafficking and intelligence, with works ranging from classification of threat intelligence risks (Santos et al., 2012) to stream reasoning technology for live detection of leaks (Parveen et al., 2013). Frequently such cyber-espionage is functional to (5) Cyber-warfare — these activities focus on operations carried out in the cyber domain to achieve an operational advantage of military significance, with a full report from the US Military intelligence (Cordesman, 2002) as a seminal work; (6) DDoS/Spam/Phishing — similarly to cyber-espionage Distributed Denial of Service (Kandula, 2005), Spam or Phishing criminal activities are connected to crimes against critical infrastructures (Setola et al., 2016); (7) cyber drug-trafficking — these activities focus on the stockade, movement, production, and reselling of illegal substances, wit early works focusing on identifying the extent and properties of the collaboration networks lying beneath (Wood, 2017).

Overall for the above crimes all have been reported in connection to software-based electronic threats, vulnerabilities, and attacks where Malware (including ransomware and similar malware aiming explicitly at financial gains), Trojans, or Botnets play an instrumental knowledge-gathering and insurgency role, with the latest works on this research stream discussing the architectural properties of malware altogether, e.g., Lakhotia and Black (2017).

4.3. *Cybercrime threat intelligence: a taxonomy for deep- and dark-web*

Beyond the previously defined taxonomy addressing the surface web, this section discusses the approaches, countermeasures, indicators for Cybercrime Threat Intelligence in the deep- and dark webs. The taxonomy in question (see Fig. 11) shares overlaps with its surface web counterpart (see Fig. 10), specifically in the criminal activity types and countermeasures thereof (see the greyed boxes with ".*" symbol).

4.3.1. *Assessment methods (METH)*

The assessment methods harnessed for the investigation in the context of deep-web and darknets are considerably different with respect to their surface internet counterparts. Data indicates a distinct use of port-scan (Gadge and Patil, 2008; Kikuchi et al., 2008) techniques as a basis for assessment, namely, detecting port activity in or around a specific host. Most recent works along these lines reported in our data set are from Neu et al. (2018) offer a glimpse of port-scan technology in the context of Software-Defined Networks (SDNs) (Sorensen, 2012) as well as Ring et al. (2018) who manage to detect port-scans at large-scale. A similar attempt to Ring et al. comes from Affinito et al. (2018), who implement a stream analysis campaign over Apache Spark to instrument for large-scale port-scans. From a higher level of abstraction, network scans (Mazel et al., 2016) are reported as the second most frequent method for online source risk assessment; network scans are defined as a procedure for identifying active hosts on a network, either for the purpose of attacking them or for network security assessment (Leckie and Ramamohanarao, 2002). Recent research in this domain shares the same aims as port-scanning research, i.e., detection and avoidance. Concerning the remainder of the approaches, a very valuable recap is offered by Liu and Fukuda (2018). More specifically, OneFlow analysis concerns analysis of large-scale traffic directed at single entry-points inside a network (Nishikaze et al., 2015; Yegneswaran et al., 2004) while Backscattering (Balkanli et al., 2015) and IP-Fragment analysis (Kim et al., 2013) concern identifying different aspects of DDoS attacks, namely, response packets to (D)DoS attacks carried out elsewhere in the Internet and attempts to defeat packet filter policies. Furthermore, small-* analysis techniques aim at establishing anomalies in network traffic reflecting a minimum amount of specific packet types (e.g., SYN, UDP, Ping) — a very valuable comparative outline of these approaches is contained in Kumar and Mittra (2014).

4.3.2. *Anonymous crawling policies (ACP)*

Considering the invasiveness of assessment methods, we were not surprised not to find many approaches to anonymous crawling and cybercriminal activity assessment. The few literature elements that do exist discuss the use of countermeasures to graph-mining (Phillips and Lee, 2009) as mechanisms to prevent the detection of cybercriminal activity assessment, e.g., by rearranging network topologies by means of software-defined networking. Similarly, Haughey et al. provide evidence for the use of traffic randomization to avoid adaptive traffic fingerprinting (Haughey et al., 2018).

4.3.3. *Risk-level parameters (SRLP)*

Risk-level parameters offered by literature in threat intelligence range from infection attempts coming from a specific source (Yannikos et al., 2018) as well as misconfigured request-response messaging patterns (Fachkha and Debbabi, 2016); in this context, the recurrent use of backscattering counts from specific sites have been reported as indicative of high-risk online sources (Fachkha, 2016). Likewise, the number of probe code instances, that is, code designed to attempt gaining access to a networked host and its files through a known or probable weak point, has been established as a proxy for high-risk online sources (Canepa and Claudel, 2013). Finally, risk analysts can use an assessment of a small-world topology condition reflecting the links and call-forwarding structure stemming from the source (Kleinberg, 2000; Narayanan and Shmatikov, 2009).

4.3.4. *Software quality parameters (SQUAL)*

Again, the literature is not conclusive in terms of software quality characteristics that can be used as a proxy for cybercriminal activity. On the one hand, the use of graph-based intelligence includes parameters such as call-forwarding occurrences (Huang et al., 2018) in the online source code as well as inferential social-networks metrics applied to call-forward graphs (Bryan Monk and Davies, 2018). On the other hand, related literature in malware detection and avoidance suggests the study of malware code (e.g., counting honeypot wrappers Bou-Harb et al., 2015; Schneider et al., 2011) to identify signature structures matching specific cybercrime (Shosha et al., 2012) as well as studying the subnetting structure in a call-forward graph and the hosts therein (Ahrend et al., 2016; Kim and Kim, 2018).

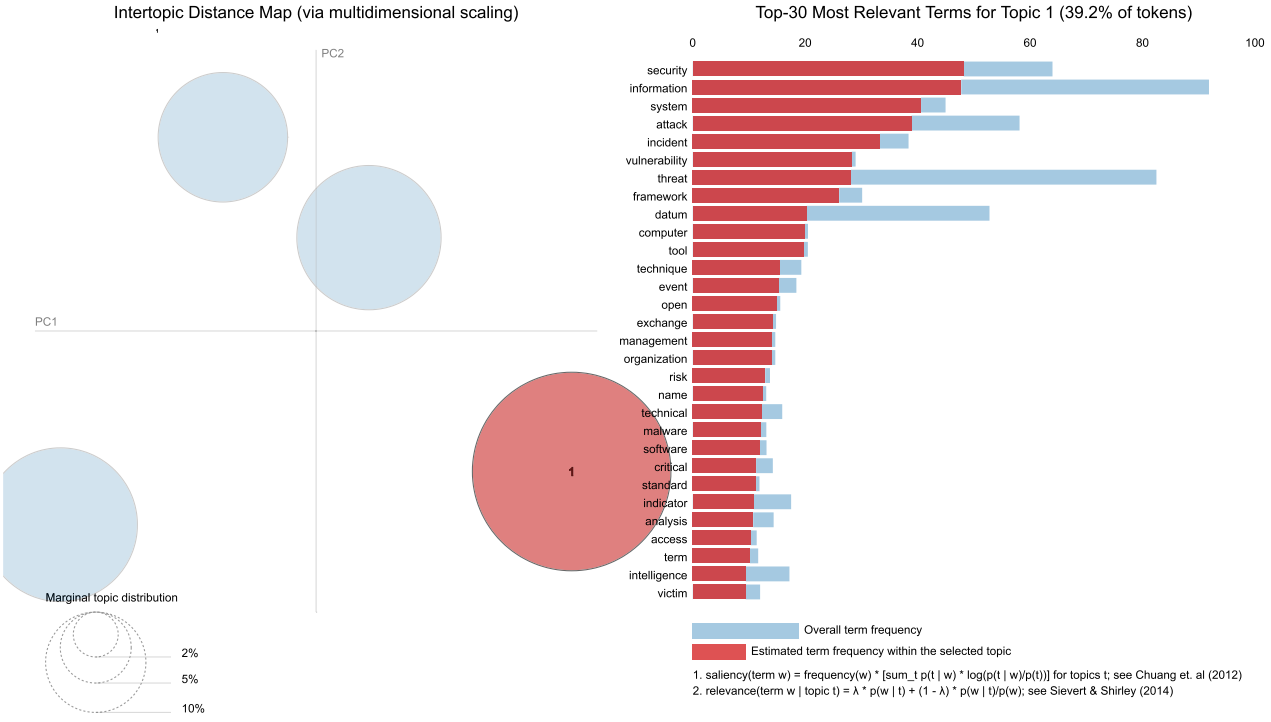


Fig. 12 – Topic modelling results of the first topic in surface web.

4.3.5. *Website-appearance parameters (WSAP)*

Finally, in terms of online sources' appearance, several parameters emerged that are germane to establishing the cybercrime risks for such online sources. Most specifically, the amount of observable (i.e., monitorable and loggable) characteristics or *observables* (Nabki et al., 2017) of the source along with the number of words employed for textual descriptions as well as typical name commonalities around the source (Narita et al., 2016; Skopik et al., 2016); individual characteristics of words and phrases (e.g., as reflected by Maximum Sentence Length Bailey et al. (2006) are also suggested as indicative (Yang et al., 2007). Beyond simplistic counts, related literature on website-appearance from deep- and darknets highlights the use of ad-sense as well as irrelevant people attributes (Wang et al., 2018) being requested for registration as primary indicators of specific cybercriminal activities (Yang et al., 2007). Specific people attributes (e.g., bitcoin accounts and transactions thereof Khelghati, 2016) are often associate to illegal-trafficking.

4.4. *Cybercrime threat indicators: topic modelling results*

To address SRQs 1–3 and 5–6, we adopted thematic coding (Buder and Creß, 2003) to elicit a baseline understanding of the state of the art. More specifically, the selected sample of articles was subject to annotation and labeling to identify themes emerging from the analyzed text. Before applying Latent Dirichlet Allocation (LDA), we pre-processed our text. The pre-processing phase of LDA was carried out after standard text-mining pre-processing to improve results by removing unnecessary information. Specifically: (1) all terms and definitions for the factors were standardized in terms of structure (i.e., definition + sample text extracted from reference papers); (2) punctuation marks and numbers were removed; (3) all letters were converted to lower case; (4) all common stop words for English grammar and syntax were removed.

After the pre-processing phase, we apply the LDA method for visualizing and interpreting topics. The method we used is the one described in Sievert and Shirley (2014) called **LDAvis** and based on the work of Chuang et al. (2012). Moreover, the paper gives instructions on reading the diagrams we plotted; however, below, continue with a small recap about how to interpret our diagrams. On the left side of our figures, we have a recap of our topics. Each of the circles represents a topic and how prevalent it is. Moreover, if the circles are overlapping each other means that those topics have common terms. Into each of these circles are sorted our terms in decreasing order of prevalence.

Our results' right panel depicts a horizontal bar chart of the most useful term to interpret the selected topic. The overlaid bars represent both the corpus-wide frequency of a given term as well as the topic-specific frequency of the term (Chuang et al., 2012; Sievert and Shirley, 2014). The λ slider allows ranking the terms according to term relevance. Moving the slider allows adjusting the rank of terms based on much discriminatory (or "relevant") are for the specific topic. We fixed the λ at 0.8 to highlight frequent terms but not exclusive, a λ equal or close to zero will highlight potentially rare but exclusive terms for the selected topic.

Here below we are now going to discuss our results of our topic modelling analysis. For each analysis for the Surface Web in Figs. 12–15 we build a table where we summarize and discuss the most relevant terms related to the cybersecurity field. We then will do the same for the analysis for the Deep-, Dark-Web Web in Figs. 16–19.

Table 2 – Topic analysis results of the first topic in surface web.

Terms	Score	
System	40	The system can receive different types of attacks. The operating system of a PC, if not properly maintained, can easily be the target of virus, worm, malware, spyware and other cyber threat attacks. In order to protect the system of a private user is important to have an antivirus and keep the system updated. In a private company is important to educate the employee to do not use distrusted applications and to use strong passwords in order to protect personal information. An attack to the system can involve a loss of personal data, a destabilization of the running processes of the system and the forward of private information to third parties.
Vulnerability	27	Techopedia defines the therm *vulnerability* as a bug in a system that make the system itself unsecure and opened to attacks. Some computer vulnerabilities include bugs, weak password, outdated operating system, OS command injection, download of pirated software (Techopedia, 2019a)).
Threat	27	We live in a hyperconnected world, half the world's population is interconnected through Internet and 125 billion of IoT devices are expected by 2030 to be connected. All this complexity along with constantly evolving nature of cybersecurity threats is leading to more breaches and cyberattack threats. Threats also known as vulnerabilities can turn into attacks on computer systems, networks, and more (Techopedia, 2019a)).
Malware	17	Malicious software (Malware) is one of the most common cyber threat attack. A Malware is any software that does harm to the system, such as a virus or spyware. There are a lot of different versions of Malware: virus, trojan, rootkit, worm, spyware and adware, all of them with different characteristics but with the same purpose. The aim of all this malicious software is to steal private information from the victim's PC, profile the habits of the victim user, use the attacked machine as a zombie for network attacks.
Software	17	As software prices increase, many users turn to installing bootleg copies, or pirated ones. According to the study in Microsoft Philippines PR Team (2017) 34% of the downloaded pirated software came bundled with malware that infect the computer once the download is complete or when the folder containing the pirated software is opened. In order to avoid any type of risk, a solution is to use a free version of the software or similar software but free or open source.

Table 3 – Topic analysis results of the second topic for surface web.

Terms	Score	
Packet	19	Packets are used in a Denial of Service (DoS) attack in order to make inaccessible services of those machines in the network. DoS attack's main targets are usually web servers of high-profile organizations such as media companies, government, trade organizations, banking, and e-commerce platforms. In general, we have two types of DoS attacks: flooding services and crashing services. The first is caused by high traffic to the servers that make the services slow down and eventually stop. In the latter case, the DoS attack exploits vulnerabilities to let crash the running services or destabilize the system.
Link	10	Also known as Url is a unique identifier used to locate a resource on the internet. However, often Url's are used to carry unaware users on distrust websites built in order to steal personal information and banking coordinates and passwords. Url's are spread through emails, gaming platforms from OSN (Online Social Networks), SMS and instant messaging platforms.

4.4.1. *Topic modelling results for surface web*

In the following tables and figures, we are going to discuss the results of our topic modelling analysis for the Surface Web. The tables contain the most relevant terms cybersecurity related of each topic. Meanwhile, the figures show the results with all the most relevant terms.

The relevant terms from our first topic and listed in Table 2 highlight the most common cybersecurity threat at the operating system (OS) level. The OS of a user is indeed the first target of hackers and criminals to steal information. If a system is not updated or protected through antivirus or firewall, it can be easily attacked by criminals. A famous vulnerability in Microsoft OS is *EternalBlue*; this vulnerability targets the Microsoft Windows Server Message Block (SMB) protocol and allows attackers to execute arbitrary code. This vulnerability is extensively used today by ransomware like WannaCry, Petya, and NotPetya. These threats infect the user's PC and encrypt the whole hard drive asking for a ransom to receive a key to decrypt all the files. Moreover, the common practice of downloading pirated software from unknown sources makes life easier for hackers, viruses, and malware.

In Tables 3 and 4 we have the list of terms for our second and third topic. Here we notice that our topic analysis indexed all the possible vehicles for any cybersecurity threat. We, indeed, have links usually used in order to launch a phishing attack. Furthermore, emails are used by the phishing attack to carry a message to trick the recipient into believing that the message is something they want or need, similar to legit advertising or like a genuine request from the personal bank or a note from someone in their company. These emails are used to push the reader to download an attachment or share

Table 4 – Topic analysis results of the third topic for surface web.

Terms	Score	
Social	12	OSN (Online Social Networks) are the new life place for many people, they use this platforms to keep in touch, to share pictures and comments, to read news. However, these OSN platforms can easily become a cybersecurity threat for every user. Third-party apps: hackers may be able to gain access through vulnerabilities in third-party apps that integrate with the big social networks. Phishing attacks: using fake promotions or through the promise of significant discounts, hackers can tempt users to click on phishing links in order to steal banking information. Identity theft: collecting the public information available OSN, hackers can turn into unaware user in order to perpetrate scam. Confidential information leak: can happen when not expert users are not able to set up privacy settings in order to protect personal information.
Website	10	A website can be the place of cyber threat attacks if the system behind it is not well updated or if the admin's passwords are not strong enough. A compromised website can host different types of cyber threat attacks. A Phishing website can steal the passwords and personal information of a user. A website can also be the target of SQL Injection Attacks to retrieve private information from the database.
Email	9	Emails are one of the main vehicles of information and data. The Identity Theft attack happens when an attacker can gain a handle on the employee's email account. The attacker can then turn into the employee's identity. Phishing Attacks is a type of social engineering attack often used to steal user data, including login credentials and credit card numbers. It occurs when an attacker, masquerading as a trusted entity, dupes a victim into opening an email. The recipient is then tricked into clicking a malicious link. Virus as an attachment to the email to install unwanted software on the PC of the user. Spam email is commonly used to deliver Trojan horses, viruses, worms, spyware, and targeted phishing attacks or to bring users on an external website to steal private and personal information.
Process	5	Processes are all the related activities (parts) inside the system that work together to make it function. A compromised process can lead to an unstable system. We have different types of attacks, viruses, worms, malware, spyware, and trojans to compromise a process. All these attacks target the system's processes to change the main functionalities and make them work for the attacker. To avoid this type of attacks, it is extremely important to have an updated system, a proper antivirus installed, and a firewall to defend the system environment.

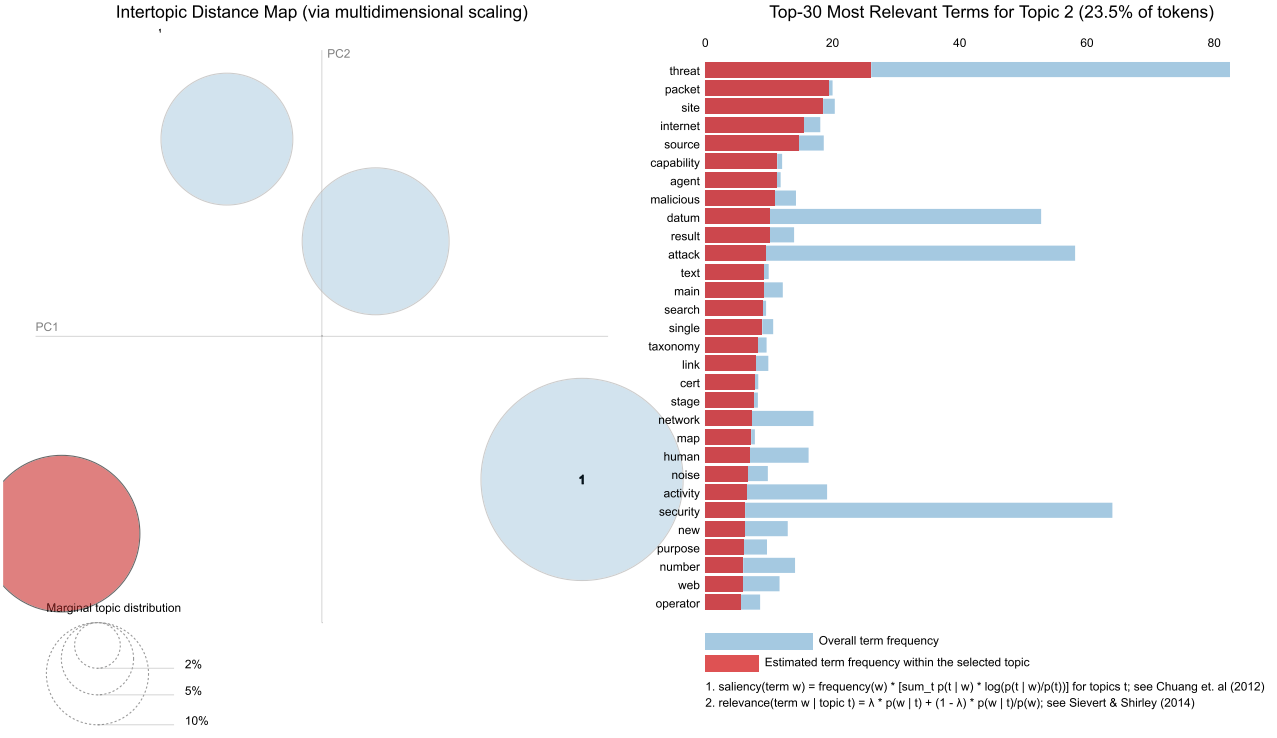


Fig. 13 – Topic modelling results of the second topic in surface web.

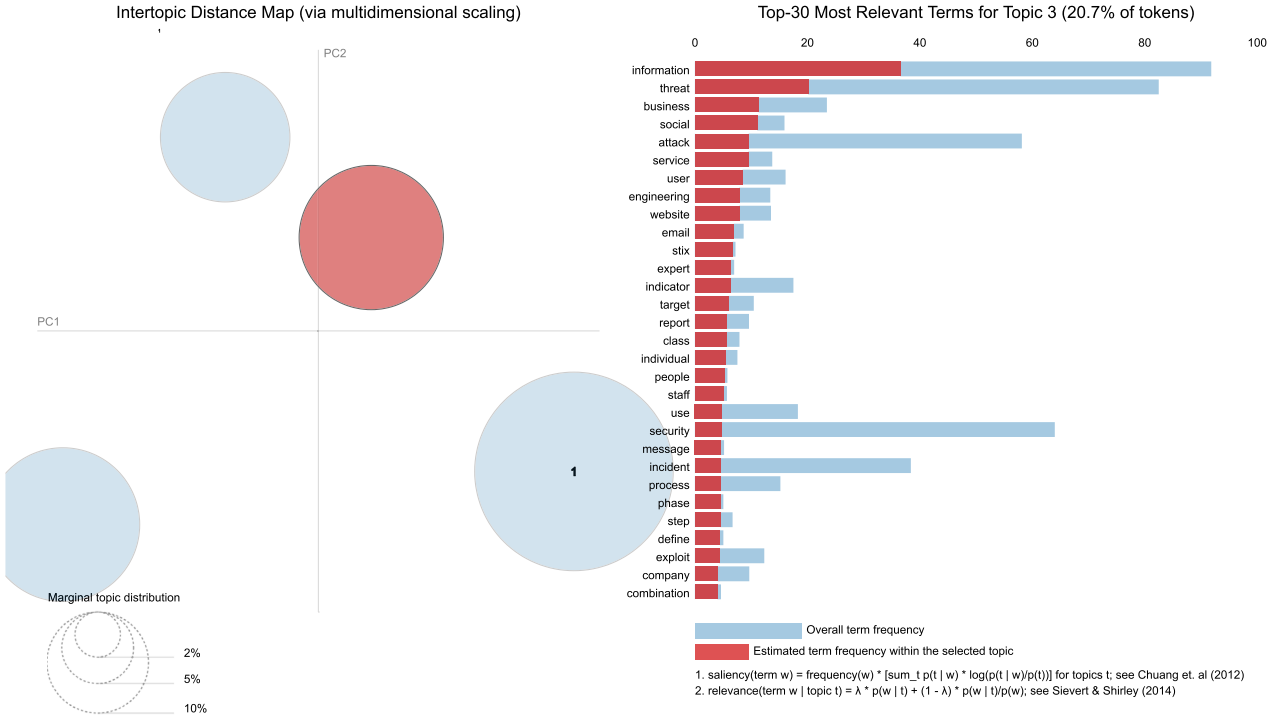


Fig. 14 – Topic modelling results of the third topic in surface web.

Table 5 – Topic analysis results of the fourth topic for surfaceweb.

Terms	Score	
IODEF	8	Incident Object Description Exchange Format defines a data representation that provides a framework for sharing information commonly exchanged by Computer Security Incident Response Teams (CSIRTs) about computer security incidents. This document describes the information model for the IODEF and provides an associated data model specified with XML Schema (Danyliw, 2016)).
CTI	4	Cyber Threat Intelligence is what cyber threat information becomes once it has been collected, evaluated in the context of its source and reliability, and analyzed through rigorous and structured tradecraft techniques by those with substantive expertise and access to all-source information (CIS).
OTX	4	Open Threat Exchange it is a platform in order to share information about threats and provides access to a global community of threat researchers and security professionals. OTX allows anyone in the security community to actively discuss, research, validate, and share the latest threat data, trends, and techniques.
GLBA	3	The Gramm-Leach-Bliley Act is also called the Financial Modernization Act of 1999. It was passed by Congress as a means of controlling ways in which financial institutions handle and deal with individuals private information.

personal credentials like passwords and bank accounts numbers. Online Social Network nowadays, platforms where users share every personal data daily. This information database can be the target of a lot of different attacks range from cyberbullying, identity theft, phishing to viruses from third-party apps, fake profiles, etc.

The last topic in Table 5 lists some standard format for computer security incident response, all of them with the idea of collaboration to decrease the risk of a cyber threat at the company level. IODEF is an object-oriented structured format used to describe computer security information for exchange between Computer Security Incident Response Teams (CSIRTs). OTX is a platform built to share information about threats and have a network of researchers and security professionals that can collaborate to handle the threat.

4.4.2. *Topic modelling results for dark and deep web*

In the following tables and figures, we are going to discuss the results of our topic modelling analysis for the Deep, Dark-Web. The tables contain the most relevant terms cybersecurity-related to each topic. Meanwhile, the figures show the results with all the most relevant terms.

In Table 6, we have terms strictly related to the Dark and Deep web. The bitcoins are, indeed, the most used cryptocurrency in order to keep transactions and illegal trafficking anonymous. We have then *Domain* term that highlights that

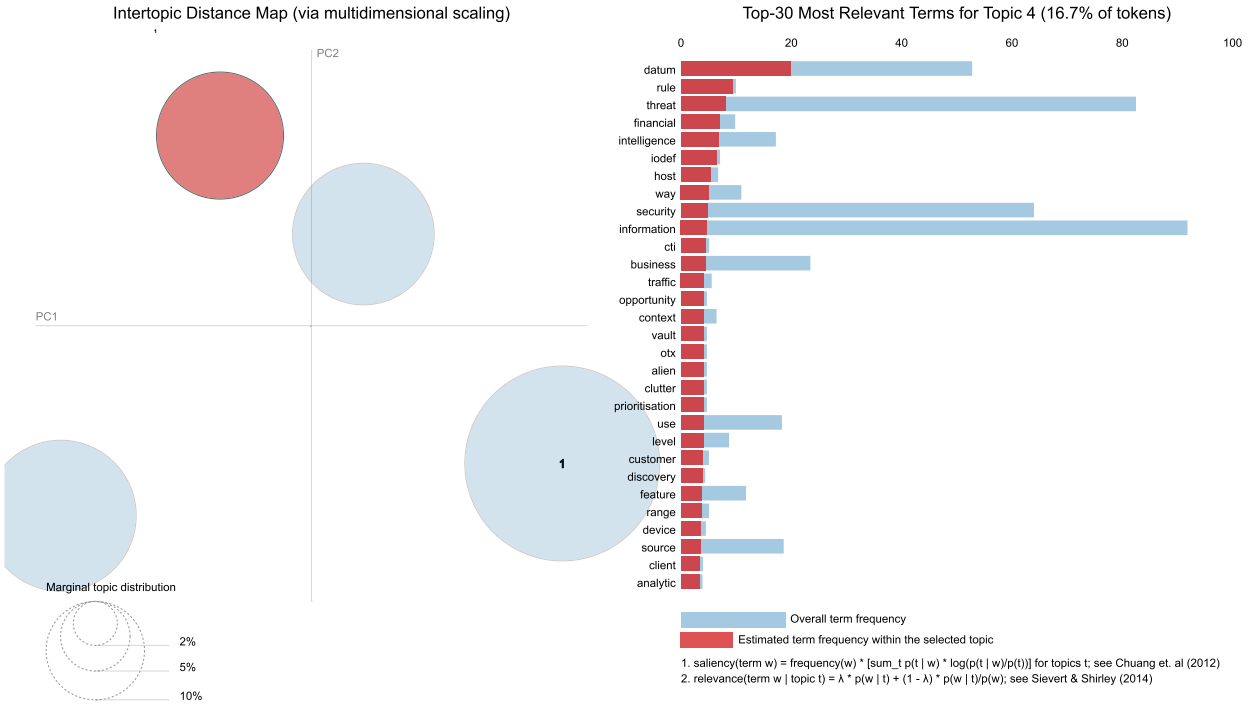


Fig. 15 – Topic modelling results of the fourth topic in surface web.

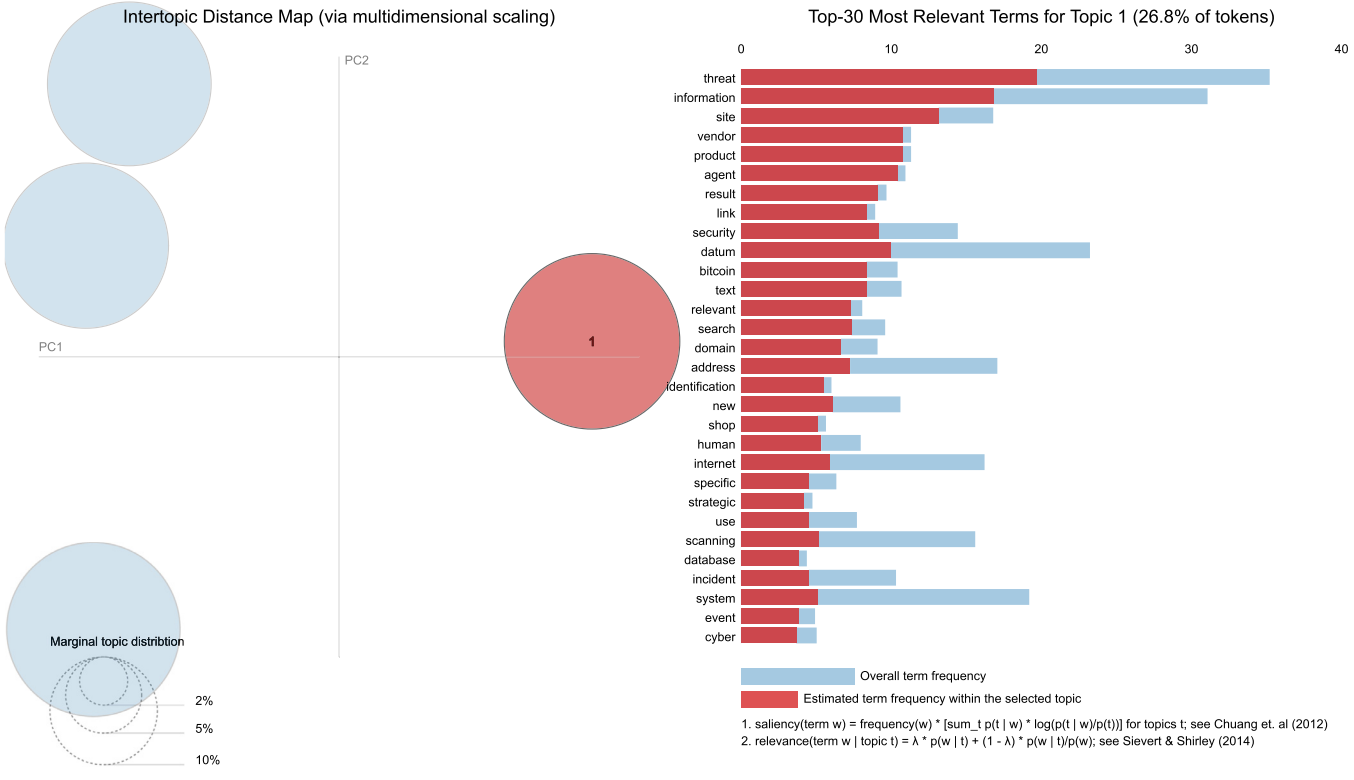


Fig. 16 – Topic modelling results of the first topic in dark and deep web.

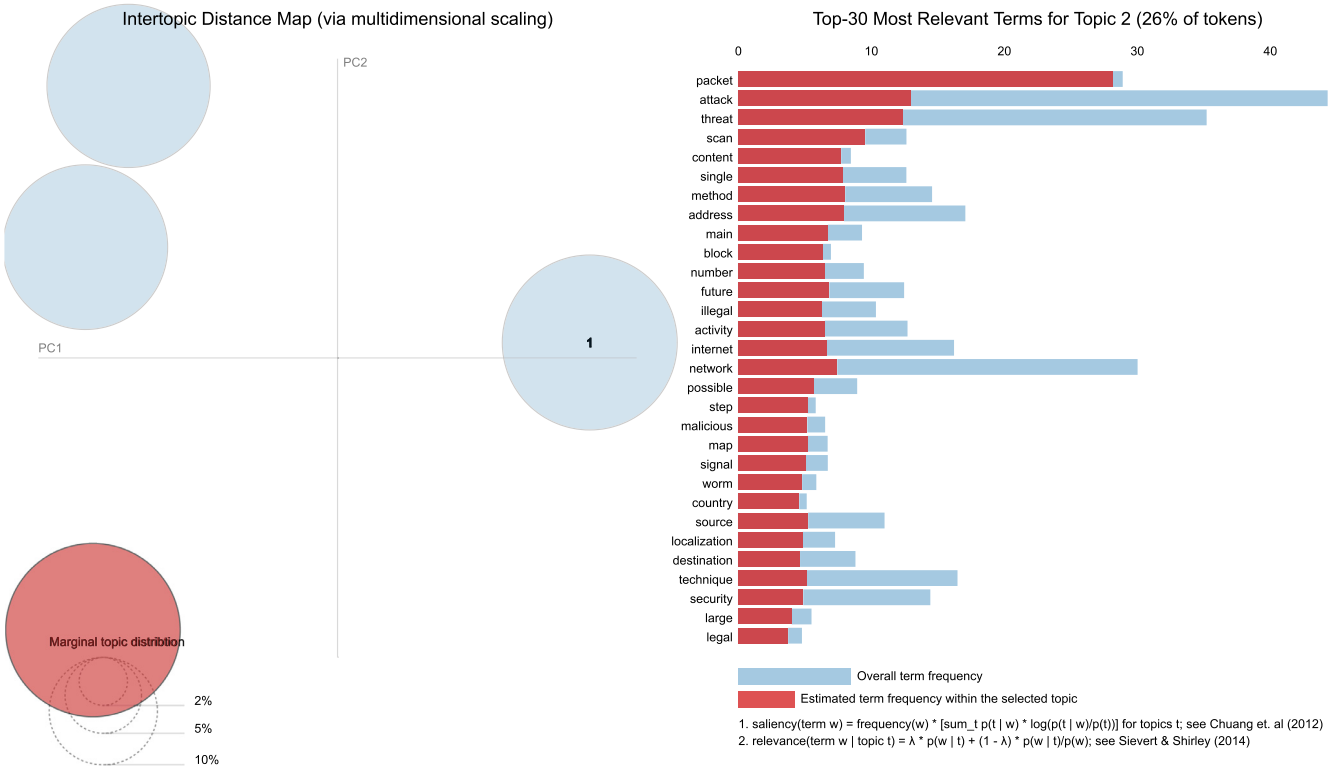


Fig. 17 – Topic modelling results of the second topic in dark and deep web.

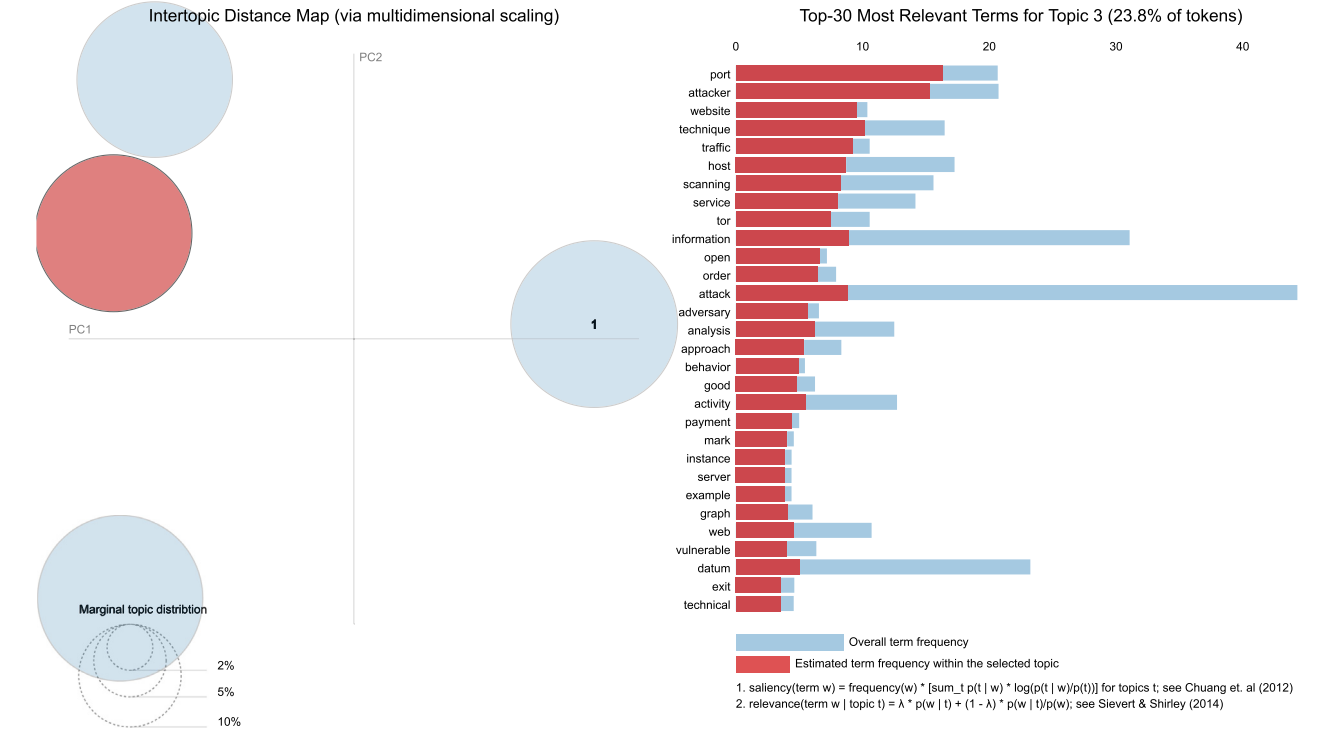


Fig. 18 – Topic modelling results of the third topic in dark and deep web.

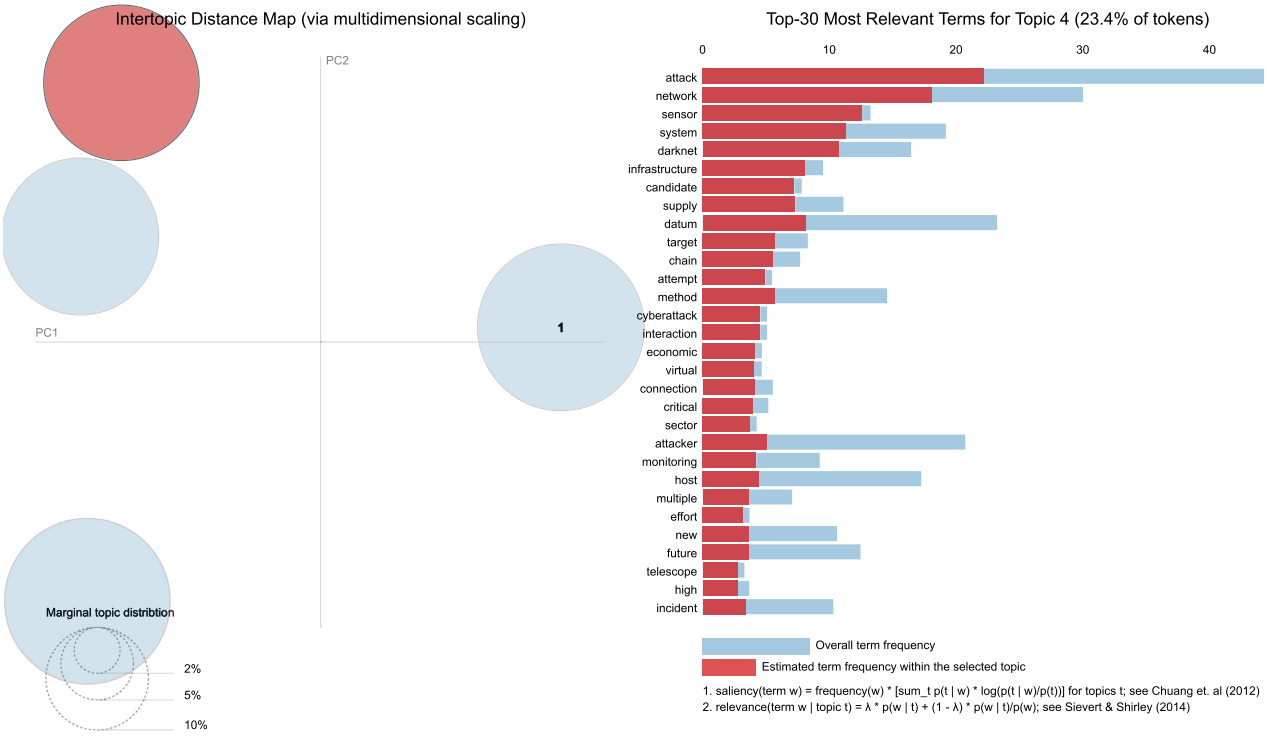


Fig. 19 – Topic modelling results of the fourth topic in dark and deep web.

Table 6 – Topic analysis results of the first topic for dark and deep web.

Terms	Score	
Bitcoin	9	In the last years, we witnessed a huge growth of this kind of currency and transactions. As a virtual currency, Bitcoins can also ensure a high level of anonymity and for that reason is widely used for illegal transactions. In the last year bitcoins have been used to ask ransom after a cyberattack, an extensive use in Dark and Deep web for illegal trafficking and the adoption of this virtual currency in all those activities that need a high level of anonymity.
Domain	8	*"Internet Domain is a unique name on the Internet. The chosen name combined with a generic top-level domain (gTLD), such as.com or.org, make up the Internet domain name"* (Pcmag Encyclopedia, 2019)).
Shop	7	Dark and Deep web are the best places where to trade illegal products. The nature of the Dark web together with the use of cryptocurrencies like Bitcoins, make possible to sell and buy every type of product from armies to drugs, counterfeit medicine, and bullets being completely anonymous.

Table 7 – Topic analysis results of the second topic for dark and deep web.

Terms	Score	
Worm	5	Worms are similar to computer viruses and works in order to alter the functionality of a system. A worm exploits the vulnerabilities of the system in order to takes advantage of file-transport or information-transport features on the system, allowing it to travel unaided. Worms like WannaCry, Petya or NotPetya are more sophisticated and are able to leverage encryption, wipers, and ransomware technologies to harm their targets.

all this transaction takes place in reserved space. Indeed, all the domains in the darknet are.onion since they are part of a different network type. Lastly, the term *Shop* point out one of the most common activities in the dark web, the trade of illegal items.

In Table 7 we highlight the term Worm. WannaCry, Petya, and NotPetya are examples of worms usually born in the darknet built by hackers or by cybercriminals. Most hackers community have their virtual space in this part of the network where they sell every type of cybersecurity threat, attack, and "services" to steal private information.

Table 8 – Topic analysis results of the third topic for dark and deep web.

Terms	Score	
Tor	9	Tor is a web browser that using the Tor network is able to anonymize your traffic protecting then your identity online (Porup, 2018)).

In Table 8, the third topic table related to the Dark and Deep web, we highlighted the term Tor. Tor is the network of servers behind the Darknet. It provides anonymity, and in order to access this network, we have the Tor Browser. Tor Browser anonymizes all the traffic routing it through the Tor network. Tor is a multi-layer proxy and connects at random to one of the publicly listed entry nodes. It then always selects at random one middle relay and finally spits out the traffic through the third node (Porup, 2018). This type of network makes it impossible to trace users and illegal activities.

In the last topic in Table 9 we have the definition of Deep web and Dark web. Unlike the surface web, the deep web does not allow search engines to crawl and index websites. Usually, the Deep web is a safe place where the content is not available on the surface web and remains private. Differently, the Dark web operates with a high degree of anonymity and hosts harmless activities and content, as well as criminal ones. What makes it possible to do business on the Dark web is Bitcoin and cryptocurrency that helps assure buyers and sellers anonymity.

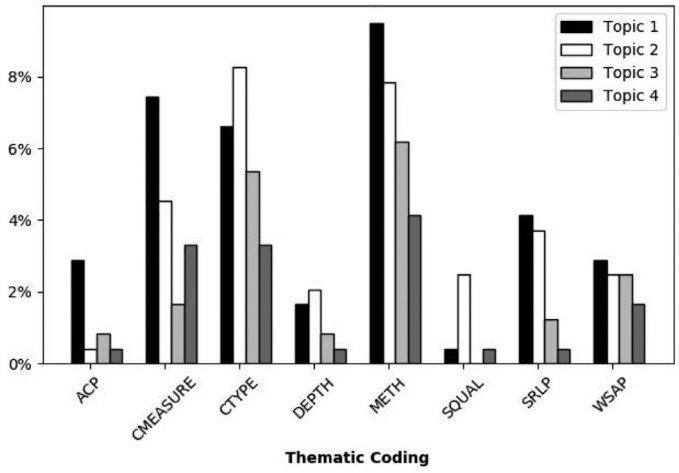
4.4.3. *Qualitative vs. quantitative insights: overlap and considerations*

In this last section, we have in Fig. 20 the distribution of our thematic codes among the different topics in Surface Web and Deep-, Dark-Web. We notice that for our thematic coding, all the subjects are covered in a Surface Web. Meanwhile, in the Deep-, Dark-Web distribution, we assert that there is a gap in scientific studies on **ACP** (Anonymous Crawling Policies), **SQUAL** (Software Quality), **WSAP** (WebSite Appearance Parameter). It means that the scientific community did not produce any study about Anonymous Crawling in a Deep and Dark-Web and neither for SQUAL and WSAP. These results can be translated into a lack of research studies from the scientific community. Clearly, the lack of researches in the field of Anonymous Crawling is strictly related to the privacy-preserving nature of the Deep and Dark Web. The Deep and Dark web is an "unauthorized" space where the information needs to be protected and secured and aims to guarantee users' anonymity. To this regard, most of the illicit web markets, forums, child-pornography platforms, and website in Deep and Dark Web are protected with rudimentary protection mechanism, unlike captcha security login, weblink redirection, and with very limited lifetime of these illicit platforms (usually websites appear and disappear in a day Sanchez and Griffin, 2019). This explains the reasons behind the difficulties faced by LEA's in cybercrime fighting in the Deep- and Dark-Web. Based on the above observations, in conjunction with our literature review study, we strongly put forward the following recommendations. Indeed, the development of an Anonymous Crawler needs, first of all, to be supported by a (more) effective VPN (Virtual Private Network) tool. Moreover, the scientific community must start leveraging technologies like WebDrivers (i.e., Selenium Webdriver, CasperJS, PhantomJS) and HTML parsing packages (i.e., BeautifulSoup, Scrapy, lxml) to extract information automatically while protecting the investigators from directly "viewing or reading" explicit, inhumane and violent content from the Deep and Dark Web. Conversely, to bypass captchas, researchers should focus on identifying, evaluating, and testing new approaches to improve fully automatic captchas' resolution to solve the dependence on manual, time-consuming and cumbersome captcha solving. On the other hand, captcha solvers exploiting technologies -like python-anticaptcha, python captcha-solver, and python Tesseract- could be put in place to be validated, and further developed and eventually instrumented with machine learning techniques to extract symbols (numbers and letters) from images or rather recognize specific images in a set of pictures. Lastly, based on the results in Fig. 20, we strongly believe that the scientific community should also further investigate and provide a set of policies (**ACP**) to crawl the Deep- and Dark-Web ensuring the anonymity of the investigator. The same lack of studies has in fact been encountered for the Software Quality parameters (**SQUAL**). There is again no effective solution to assess software quality from the Deep- and Dark-Web from a scientific perspective. This lack of studies leaves the Deep- and Dark-Web as a dark space where everything could be considered a threat. Lastly, the scientific community did not develop sufficient understanding about the WebSite Appearance Parameter (**WSAP**). As already discussed in paragraph 4.3.5, most of the time, the parameters used are those observable, like the length of the textual description, the number of words, and the use of people attributes like bitcoin accounts and transactions. Unfortunately however, these parameters are not enough to let the LEAs assess the quality of a website in the Deep- and Dark-Web making the investigation phase more time-consuming and less accurate. In conclusion, we assert that scientific community get more (direct) feedback from LEAs, and bypass the lack of works in **SQUAL** and **WSAP**. From a technical point of view, we truly believe that the development of Anonymous Crawling techniques could enhance and develop new parameters for WebSite Appearance and Software Quality. In the meanwhile, a possible solution could be to start identifying, evaluating, and testing technologies like Google Safe Browsing API,[15] PhishTank API,[16] VirusTotal API,[17] Quttera API,[18] Sucuri API,[19] GreyNoise API,[20] URLScan API,[21] Cloudflare API,[22] Shodan API,[23] Metasploit API,[24] AlienVault

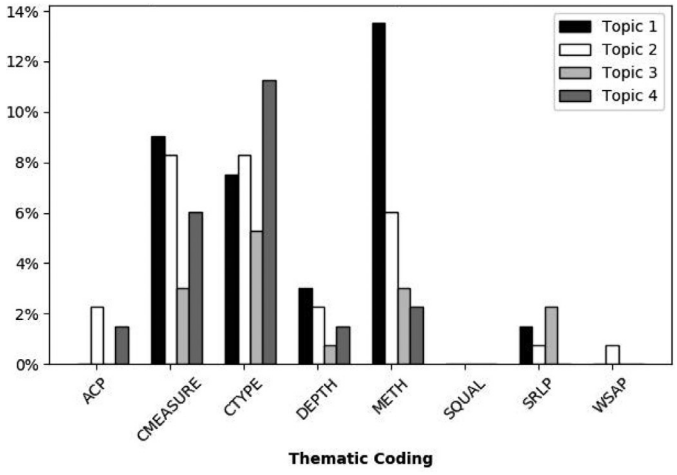
[15] https://developers.google.com/safe-browsing/v4/get-started
[16] https://phishtank.org/api_info.php
[17] https://developers.virustotal.com/reference
[18] https://quttera.com/quttera-web-malware-scanner-api
[19] https://docs.sucuri.net/website-monitoring/scanning-api/
[20] https://github.com/GreyNoise-Intelligence/api.greynoise.io
[21] https://urlscan.io/about-api/
[22] https://api.cloudflare.com/
[23] https://developer.shodan.io/
[24] https://docs.rapid7.com/metasploit/rest-api/

Table 9 – Topic analysis results of the fourth topic for dark and deep web.

Terms	Score	
Darknet	11	The internet is composed of three layers: the surface web, the deep web, and the dark web. On top we have the surface web, here we have web pages indexed by all the search engines such as Google, Bing, or DuckDuckGo. In the deep web, we have those web pages that are not indexed by the search engines. Therefore these pages and websites are hidden to the surface web side and usually can be accessed through passwords and authorization. Then the lowest level of the web is characterized by the dark web. This part of the network is untraceable online together with its activities. The dark web cannot be found using common search engines, and you need to use specific software and configurations. The dark web's main purpose is to keep their illegal web activities hidden (Hale, 2018)).



(a) Surface Web.



(b) Deep-, Dark-Web.

Fig. 20 – The two figures describe the thematic coding distribution in surface web and deep-dark web topic analysis.

API,[25] SecurityTrails Data Security API,[26] on offline website dumps crawled from the Deep and Dark Web to start collecting information on the most common WebSite Appearance and Software Quality parameters used. Overall, this analysis thus suggests that further studies in these directions are needed to fill the widening scientific gap among the Surface Web and the Deep- and Dark-Web.

5. Discussion

This section explicitly addresses our research questions. For the sake of space, specifically related research questions (e.g., SRQ1 and SRQ5 as well as SRQ2 and SRQ3) are collapsed into

[25] https://cybersecurity.att.com/documentation/api/alienvault-apis.htm
[26] https://docs.securitytrails.com/docs/overview

one coherent answer, substantiating that answer with either the descriptive statistics or other results from Section 4.

5.1. *What online depth levels are assessed, and to what extent?*

This research question was originally aimed to assess the extent to which state of the art has targeted exclusively surface, deep, or dark webs, respectively, and with which technique. Our data indicate a wide array of methods spanning all three levels of depth with sensible overlaps in between, e.g., the usage of attack graphs and social-networks analysis of the involved networks, as reported in both the taxonomies distilled and reported in the previous pages. In summary:

Finding 1. Deep and darkweb cyber threat engineering and management have predicated much on network-based analysis as well as low-level artifact mining (e.g., packet mining, code analysis, etc.). More higher-order and multi-vocal data is remaining unused and deserves further attention.

5.2. *What degrees of anonymity exist for web-crawling?*

This research question aimed to identify the mechanisms and approaches to ensure the inquirers' anonymity while crawling or analyzing the shallow, deep, and dark network levels. Our data is rather inconclusive since many of the approaches (e.g., template avoidance) are consistent with specific investigation types. Moreover, according to both taxonomies for surface and deep/darkweb, the generalization of 'degrees' of anonymity is impossible at this stage, and further research is needed in this direction. In summary:

Finding 2. There exists no conclusive crawling/analysis anonymity procedure in state of the art; this avenue is open for further research opportunity and is urgently in need of addressing by major Law-enforcement agencies across the EU.

Furthermore, concerning the subsequent research question, namely, "What policies exist to vary the degrees of anonymity?", there exist relations between the several approaches we did find (e.g., botnets in conjunction with sandboxing as observed in Lauinger et al. (2010b) but there exists no systematic approach to anonymous investigation to date.

5.3. *What website features are most indicative of cyber threats?*

This research question aimed to identify the recurrent characteristics and observable features that an online source may be exposed that can be used proactively to identify and profile the criminal activity being perpetrated therein. The data indicate that the indicators are quite varied; on the one hand, surface web literature points at using software code features (e.g., responsiveness) as well as their endurance over time (e.g., quality parameters fluctuation). On the other hand, deep and darkweb investigation seems to predilect low-level artifacts, e.g., maximum sentence length counts and similar devices to enact threat engineering. In summary:

Finding 3. Surface, web analysis literature, predilects software code features over appearance metrics for online source risk assessment; conversely, deep and dark web analysis literature seems to predilect appearance features, e.g., website text content mining. Little to no cross-fertilization between the two fields has been investigated so far and may require further attention.

5.4. *What risk assessment techniques exist?*

This research question aimed at identifying methods and techniques available for risk assessment in surface, deep, and dark web. The results give us an insightful overview of what types of risks need to be assessed, identified, and mitigated. Moreover, the Topic Analysis results for the fourth topic from the Surface Web (Fig. 15 and Table 5) provides some useful techniques for incident response and format exchange for sharing cyberattacks information among companies.

Finding 4. Surface, web analysis literature, showed to be affected mainly by Malware attacks and leverage of Software and Operating System vulnerabilities (Table 2). While, from Table 4, we can assert that social attacks occur through the use of Social Media Platform due to the usage of third party applications and scam messages, malicious websites thought malicious code, and emails through phishing links. Lastly, from Table 5, we infer standards for computer security incident response and collaboration techniques to decrease the risk of a cyber threat at a company level. Hence, to be useful and proactive, risks assessment techniques need to keep in consideration these type of attacks and the related environment (users cybersecurity knowledge, operating system in use, type of software and update frequency, social media platforms, the possibility of introduction of an external carrier for malware).

5.5. *Findings relevance and concrete recommendations*

Here, we now elaborate on the relevance of the insights of Section 4.4.3 and summarized in Fig. 20 - and discuss and explore how academics can use our results to shape new research paths to fill the gaps found in this literature review. On the other side, we elaborate how practitioners may use our research to improve cybersecurity and find new methods and techniques to apply. In **Finding 1.** our research highlighted how the most used parameters to predict cyber attacks remain packet mining and code analysis. However, from a research point of view, we want to encourage researchers to move forwards on a new type of analysis. Specifically, we believe that machine learning (ML) and artificial intelligence (AI) algorithms can play a key role in reducing cyberattacks. An example could be to exploit AI to continually monitoring forums from the Dark Web to predict possible attacks. Conversely, ML algorithms can focus on Dark Web markets to track new worms and viruses and plan a strategy to reduce the consequences of a cyber threat. However, ML and AI algorithms need further research to tackle cyber threats, monitor networks at runtime, reduce false positives, and implement new algorithms for upcoming threats. Moreover, the newly emerging field, namely Cyber Insurance, needs to be further developed and better consolidated in real-world settings to allow small and medium companies to transfer the risks to providers better equipped in fighting cyberattacks. In **Finding 2.** our literature review underlined the need for better crawling techniques to ensure law enforcement agencies' privacy and security. Our literature review did not find any relevant

work that implemented procedures to follow to guarantee investigators' anonymity. Hence, we truly believe that this field needs further attention from academics. With respect to this, major attention should be placed to defining the procedures to follow on how to crawl the deep and dark web privately. From a technical vantage point this implies that deep and dark web crawlers need to better deal with protection measures like captchas and mirror-link generation. Some technical solutions have been already discussed in paragraph 4.4.3. **Finding 3.** the literature review revealed a substantial gap among which website features are most indicative of cyber threats in Dark web in comparison with those found at Surface web; features at the Surface web are much better understood and complete in nature. This can be explained considering that scams and frauds through phishing pages are more frequent on the Surface web. In contrast, on the Dark web, scams and frauds are typically related to selling fake products. Our literature review highlighted the need for new code feature parameters in conjunction with appearance features to improve the recognition of malicious parameters hidden behind Dark Web portals and applications. This last **Finding 4.** is mainly relevant from an industrial perspective. Indeed, the research community needs to work more closely with the industry to define new standards and techniques to decrease the risk of cyber threats. Cyber Insurance represents an emerging field; however, it likewise eagerly needs new standards and techniques to share cyber risks information among companies while keeping sensitive private data. Moreover, there is a need to create awareness among companies and employees regarding web risks and manage personal data like passwords, emails, personal information, and confidential company information.

6. Research roadmap

As a stepping stone for a future research roadmap we herein collect and analyse the latest research contributions and insights in the field of threat intelligence. For this purpose, and to ascertain consistency, we decided to re-run our analytical queries from Section 3, contrasting them against older works, and detecting trends.

As already highlighted in Fig. 20, the new findings fall back into the METH category both for the Surface and Dark web. We can assert that the METH category appears to be one of the most prolific research fields. More specifically, we found five new papers proposing a new type of assessment method for the surface web. Interestingly, four of them are based on a machine learning approach to detect and predict an attack, while only one suggests the usage of a VPN (virtual private network) to minimize the potential impact of cyber-attack. In the Dark Web discourse, we found three publications that likewise fall in the METH category. It is interesting to notice how the three publications propose assessment methods to anticipate, predict, and mitigate a possible attack launched from the Dark web. All of them consider phishing, social attack, worms, DDOS, and botnets monitoring forums through Natural Language Processing (NLP) ML and leveraging Open Source INTelligence. Ergo, we observed from the most recent literature, and substantiated our claim that cyber threat intelligence increasingly considers to use machine learning/AI to predict a possible attack.

Moreover, as explained, we have witnessed the rise a new emerging field mainly propelled by industry to transfer a cybersecurity risk from the SMEs to insurers: Cyber Insurance (CI). Whilst Cyber Insurance topic still lingers in its infancy, it is footprint is growing at a fast pace. In particular, the CI provides a cyber risk transfer in the form of policies and implements cyberattack prevention and mitigation services to help companies of all sizes, but in particular, SMEs. Hence, the CI provides common coverage and services to train the customers to respond more intelligently during a cyberattack. Indeed, the demand for CI policies from medium and small companies grows quickly as they do not harness dedicated teams to deal with cyber risk or provide adequate protection. As drawn from analysis of literature, there emerge two critical topics in CI research that need further investigation. Firstly, more research is needed to clearly define which insurance policies address which cyber risks and events. Secondly, more research emphasis should be placed on the standardization and simplification of cyber insurance language. Lastly, research should be conducted on how to demonstrate that cyber insurance can add actual value to organizations. In the CI, the small companies transfer the cybersecurity issues to the insurance companies. Hence, the insurance companies become the new target of our literature review. The problems of methods, techniques, and indicators to fight cyber risks are then transposed to the insurance companies that can use our research study as guidelines.

From 2006 till today, the cyber threat intelligence topic has been subject to exponential growth. Over time, methods, and techniques to fight cybercrime improved accuracy, efficiency, and reaction-time against cyberattacks. At the beginning solutions were mostly facing how to recover **after** a cyberattack. Typically, they suggested using backup mechanisms adopting network rules to make the hacker's job a bit more complicated. Since recently, we have begun to face an evolution of techniques to *anticipate* and *mitigate* cyberattacks. Nowadays, techniques are focusing on how to predict and avoid cyberattacks. Moreover, the huge concern about cyber threats imposed on the industry is demanding a solution to share information about attacks while keeping sensitive data private. Hence, among the literature, we observed a growing amount of studies concerning new methods and standards to provide private channels to share cyber threats information.

Grounded on our literature review, we assert that the current main research focus is to improve the efficacy of the prediction of a cyberattack monitoring of web sources in the short term. Specifically, most literature tries to develop and validate both software quality parameters to decide whether sources are reliable and trustworthy, and, website appearance parameters to predict and recognize malicious actions from a web platform. In addition, they introduce a new breed of assessment methods to evaluate the ability to withstand cyberattacks. Notably, during the last years, we have faced an incremental growth a machine learning to anticipate a cyberattack monitoring the dark web, forums, and dark markets. In the near future, we expect a significant increase of accuracy from machine learning techniques in forecast an attack, to better

tune protections against cyber threat risks in a reduced time span minimizing or even avoiding damages.

6.1. *Recap and conclusion*

This paper provides a Systematic Multi-Vocal Literature Review on the methods, indicators, approaches, and techniques previously explored for the purpose of cybercrime threat intelligence, namely, the act of gathering information over, predicting, avoiding, or prosecuting cyber-criminal activities in the surface-, deep-, and dark-webs. More specifically, the attained results provide an overview of state of the art over (a) what online depth levels are assessed and to what extent; (b) what degrees of anonymity exist for web-crawling; (c) what policies exist to vary the degrees of anonymity; (d) what website features are most indicative of cyber threats; (e) what risk assessment techniques exist. We conclude that the extant literature has concentrated on experimentation and experience reporting over several, often disconnected and isolated parts of the phenomenon with no single integrated solution but rather with often stovepiped solutions with little to no continuity between the surface and deep-/darkweb analysis and experimental synthesis. Overall, our data, results, and discussions support the road ahead outlined below.

6.2. *The road ahead*

First, there is a distinct gap between the grey literature—which mainly discusses reported vulnerabilities as well as organizational/economic/financial consequences of being targeted by cybercriminal activity—and the white research literature—which mainly focuses on offering scattered non-definitive attempts at predicting, avoiding, or protecting against specific criminal-activity types. To address this gap, we discussed our results and the limitations therein; our discussion offers a preliminary formulation of a holistic metric to assess the risk-level that any given online source may be theatre to online criminal activity.

Second, no single community encapsulates cyber crime-fighting software, tools, approaches, and techniques. Instead, these techniques or their relevant related work is scattered across as many as 30+ domain-specific communities (e.g., software security, data privacy, software engineering, distributed computing, artificial intelligence, and more). In discussing this observation, we offered descriptive statistics over our sample in the hope of pointing community leaders in the right direction while fostering cross-fertilization or community-building.

Third, there is no one definitive solution towards assisting law-enforcement agencies in their cyber crime-fighting activity. A holistic integration effort is advised.

Forth, in Fig. 20 we showed the gap and the distribution of our thematic codes among Surface, Deep, and Dark-Web. Clearly, most of the "solutions" from white and grey literature have been developed for the Surface Web. Methods and techniques related to several of our thematic codes for the Deep and Dark Web evidently were much less investigated. This can be easily explained due to their intrinsic private nature.

Fifth, the literature review did not delivery sufficient evidence in literature to reply to SRQ4. This is due to the fact that multiple features can be used to predict cyber threats. Moreover, new research for studies from 2020 showed evidence how the research is shifting to NLP and ML. Hence, at the moment there is a clear lack of studies regarding the most indicative feature to be used to predict a cyber threat from a website.

Furthermore, from our topic modelling analysis, it is possible to imply the main topics addressed by academics and practitioners. This analysis creates a baseline for LEAs in order to better understand what types of research and practical activities are carried out by the scientific community. Moreover, having this overview of the attacks addressed by academics and practitioners, LEAs are enabled to bridge together the research community and the real scenarios proposing new research fields based on their knowledge about cybercrime. This manuscript's contributions can then be used as a reference manual for them to enrich their knowledge in the required direction. To conclude, this SLR and its results have been built to give practitioners, LEAs, and academics an overview. For the practitioners and LEAs, this SLR aims to be a starting point for the investigations and highlight methods and techniques available in the literature to fight cybercrime. Conversely, for academics, this SLR has the purpose of underlining new research paths, in which it is essential to start investigating to research novelty solutions to fight cybercrime.

In the future, practitioners and researchers should strive to address the above shortcomings even further, focusing around:

(1) providing a holistic tool to aid law-enforcers in the combat against and prosecution of online criminal activity;
(2) fostering a data-driven, cyber crime-fighting practitioners community;
(3) most immediately, building tools for large-scale online data source risk-assessment of criminal activity.

We plan to conduct and refine the above activities in direct synergy with the law-enforcement practitioners with whom we have been collaborating in this work scope.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

Acknowledgements

The work is supported by the EU H2020 framework programme, grant "ANITA" under grant no. 787061 and grant "PRoTECT" under grant Nno. 815356.

Annex I

A1. *Terms and definitions*

Table 10 lists all the terms and the related definitions used in this study. The table provides on the first column *Terms* the list of those terms considered more technical and more

Table 10 – Definition of the terms from the survey paper.

Terms	Definition
Cyber Crime	In (Techopedia, 2019a) the term Cybercrime is defined as a crime where a computer is the object of the crime as hacking, phishing or spamming attack. Alternatively a tool is used in order to commit an offense like child pornography or hate crimes.
Surface Web	Surface web is the web visible to all users using internet. The websites in the surface web is indexed by search engines like Google, Bing or DuckDuckGo (League, 2018).
Deep Web	Deep web is a private web which is not visible to all the users. The deep web consist of websites which are not indexed by search engines but can be accessed through services like VPN (Virtual Private Network) and Tor Browser (League, 2018).
Dark Web	All criminal activities like drugs dealing, killing humans etc. act upon on dark web. The user can access it only using Tor Browser services (League, 2018).
Threat Intelligence	Is information an organization uses to understand the threats that have, will, or are currently targeting the organization. The primary purpose of threat intelligence is helping organizations understand the risks of the most common and severe external threats (Nassiri, 2018).
Open Source Intelligence (OSINT)	Is the knowledge gained from processing and analyzing public data sources such as broadcast TV and radio, social media, and websites. These sources provide data in text, video, image, and audio formats (osint.it, 2015).
Crawler	A crawler is a program that visits websites and reads their pages and other information in order to create entries or retrieve data (Wisegeek, 2019).
Malware	Or "malicious software", is any malicious program or code that is harmful to any type of operating systems. The main purpose of a malware is to invade, damage, or disable computers, computer systems, networks, tablets, and mobile devices, by taking control over the operations of the device and the exchanged messages (Malwarebytes.com, 2019).
Distributed Denial of Service (DDoS)	A distributed denial-of-service (DDoS) attack consist of multiple compromised computer systems that attack together all at once a target like a server or website or other network resource. The attack causes a denial of service for users of the targeted resource. The flood of incoming messages, the high amount of connection requests or the malformed packets to the target system forces it to slow down or even crash and shut down, thereby denying service to legitimate users or systems (Cloudflare.com, 2019).
Watering Hole Attack	A watering hole attack is a security exploit in which the attacker seeks to compromise a specific group of end users by infecting websites that members of the group are known to visit. The goal is to infect a targeted user's computer and gain access to the network at the target's place of employment (Techtarget.com, 2019b).
Spoofing	Is a fraudulent practice in which a malicious party impersonates someone else, usually impersonates another device or a user on the network. The communication is then sent behind this disguised source that is well known to the receiver. Spoofing is usually prevalent in those type of communication mechanisms that lack a high level of security (Techopedia, 2019d).
Honeypot	A honeypot is a decoy computer system for trapping hackers or tracking unconventional or new hacking methods. Honeypots are designed to purposely engage and deceive hackers and identify malicious activities performed over the Internet (Techopedia, 2019c).
Insider Threat	An insider threat is a security incident that originates within the targeted organization. Such threats are usually attributed to employees or former employees, but may also arise from third parties, including contractors, temporary workers or customers. Anyone who has insider knowledge and/or access to the organization's confidential data, IT, or network resources could be considered a potential insider threat (Techtarget.com, 2019a).
Man-in-the-Middle Attack (MITM)	A man-in-the-middle (MITM) attack is a way to eavesdrop the communication between two users. In the attack a third unauthorized party is able to monitor, capture and modify the communication (Symantec employee, 2019).
Hacktivism	Hacktivism is the act of hacking a website or computer network in an effort to convey a social or political message. The person who carries out the act of hacktivism is known as a hacktivist (Techopedia, 2019b).

cybersecurity/cybercrime-related. In the second column *Definitions* we provide a short explanation of the terms related to the cybersecurity environment. We provide this table to help the reader understand the whole work better; indeed, some of the listed technical words are used in our study. Meanwhile, other terms could be useful to have a better background of the cybersecurity problem we are discussing.

REFERENCES

Affinito, A., Botta, A., Garofalo, M., Ventre, G., 2018. Detecting port and net scan using apache spark. CoRR abs/1806.11047.

Agrawal, A., Fu, W., Menzies, T., 2016. What is wrong with topic modeling? (and how to fix it using search-based se). CoRR abs/1608.08176

Ahrend JM, Jirotka M, Jones K. On the collaborative practices of cyber threat intelligence analysts to develop and utilize tacit threat and defence knowledge. In: CyberSA. IEEE; 2016. p. 1–10.

Almukaynizi M, Paliath V, Shah M, Shah M, Shakarian P. Finding cryptocurrency attack indicators using temporal logic and darkweb data. In: ISI. IEEE; 2018. p. 91–3.

Apecechea GI, Eisenbarth T, Sunar B. S$A: a shared cache attack that works across cores and defies VM sandboxing - and its application to AES. In: IEEE Symposium on Security and Privacy. IEEE Computer Society; 2015. p. 591–604.

Baer WS, Parkinson A. Cyberinsurance in it security management. IEEE Secur. Privacy 2007;5(3):50–6.

Bailey MJ, Cooke E, Jahanian F, Myrick A, Sinha S. Practical darknet measurement. In: 2006 40th Annual Conference on Information Sciences and Systems; 2006. p. 1496–501.

Balkanli E, Zincir-Heywood AN, Heywood MI. Feature selection for robust backscatter DDoSdetection. In: Kanhere S, Tölle J, Cherkaoui S, editors. In: LCN Workshops. IEEE Computer Society; 2015. p. 611–18.

Beckers K, Krautsevich L, Yautsiukhin A. Analysis of social

engineering threats with attack graphs. In: García-Alfaro J, Herrera-Joancomartí J, Lupu E, Posegga J, Aldini A, Martinelli F, Suri N, editors. In: DPM/SETOP/QASA. Springer; 2014. p. 216–32.
Bilge L, Dumitras T. Investigating zero-day attacks. Mag. USENIX SAGE 2013;38(4).
Blackwell C. A security architecture to protect against the insider threat from damage, fraud and theft. Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies. ACM, 2009.
Bou-Harb E, Debbabi M, Assi C. A time series approach for inferring orchestrated probing campaigns by analyzing darknet traffic. In: ARES. IEEE Computer Society; 2015. p. 180–5.
Bouyahia T, Idrees MS, Cuppens-Boulahia N, Cuppens F, Autrel F. Metric for security activities assisted by argumentative logic. In: García-Alfaro J, Herrera-Joancomartí J, Lupu E, Posegga J, Aldini A, Martinelli F, Suri N, editors. In: DPM/SETOP/QASA. Springer; 2014. p. 183–97.
Bryan Monk RF, Mitchell J, Davies G. Uncovering Tor: an examination of the network structure. Secur. Commun. Netw 2018;Article ID 4231326,(12).
Buder J, Creß U. Manual or electronic? The role of coding in qualitative data analysis. Educ. Res. 2003;45(2):143–54.
Caballero J. Understanding the role of malware in cybercrime. ERCIM News 2012;2012(90).
Canepa ES, Claudel CG. A framework for privacy and security analysis of probe-based traffic information systems. In: 2nd ACM International Conference on High Confidence Networked Systems (Part of CPS Week), HiCoNS 2013, Philadelphia, PA, USA, April 9–11, 2013; 2013. p. 25–32.
Chang J, Venkatasubramanian KK, West AG, Lee I. Analyzing and defending against web-based malware. ACM Comput. Surv. 2013;45(4) 49:1–49:35.
Chang SE, Lin C-S. Exploring organizational culture for information security management. Ind. Manag. Data Syst. 2007;107(3):438–58.
Chen AH, Robinson KJ, Siems TF. The wealth effects from a subordinated debt policy: evidence from passage of the Gramm-Leach-Bliley act. Rev. Financ. Econ. 2004;13(1-2):103–19.
Chuang J, Manning CD, Heer J. Termite: visualization techniques for assessing textual topic models. In: Proceedings of the International Working Conference on Advanced Visual Interfaces; 2012. p. 74–7.
CIS - Center for Internet Security,. What is cyber threat intelligence?
Clarke, V., Braun, V., 2013. Successful Qualitative Research: A Practical Guide for Beginners. SAGE Publications Ltd.
Cloudflare.com, 2019. What is a DDoS attack?
Coppolino L, D'Antonio S, Formicola V, Romano L. Real-time security & dependability monitoring: make it a bundle. In: ICCST. IEEE; 2014. p. 1–6.
Cordesman AH. Cyber-Threats, Information Warfare, and Critical Infrastructure Protection: defending the U.S. Homeland. Westport, CT: Praeger; 2002.
Danforth M. LISA. In: Limoncelli TA, Hughes D, editors. WCIS: a prototype for detecting zero-day attacks in web server requests. USENIX Association; 2011.
Danyliw, R., 2016. The Incident Object Description Exchange Format Version 2.
Elstob CM. The Simulation of Criminal Detection Activity. University of Surrey, Guildford, UK; 1974. Ph.D. thesis. British Library, EThOS
Fachkha, C., 2016. Security monitoring of the cyber space. CoRR abs/1608.01468.
Fachkha C, Debbabi M. Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization. IEEE Commun. Surv. Tutor. 2016;18(2):1197–227. doi:10.1109/COMST.2015.2497690.
. Grey Literature in Library and Information Studies. In: Farace D, Schöpfel J, editors. K.G. Saur; 2010.
Fernández Vázquez D, Pastor Acosta O, Spirito C, Brown S, Reid E. Conceptual framework for cyber defense information sharing within trust relationships. In: 2012 4th International Conference on Cyber Conflict (CYCON 2012); 2012. p. 1–17.
Franklin NBS. A comparison study of open source penetration testing tools. Int. J. Trend Sci. Res. Dev. 2018;2(4):2595–7.
Gabriel A, Schleiner S, Brauner F, Steyer F, Gellenbeck V, Mudimu OA. ISCRAM. In: Comes T, Bnaben F, Hanachi C, Lauras M, Montarnal A, editors. Process modelling of physical and cyber terrorist attacks on networks of public transportation infrastructure. ISCRAM Association; 2017.
Gadge J, Patil AA. Port scan detection. In: ICON. IEEE; 2008. p. 1–6.
García-Alfaro J, Herrera-Joancomartí J, Lupu E, Posegga J, Aldini A, Martinelli F, Suri N. In: 9th International Workshop, DPM 2014, 7th International Workshop, SETOP 2014, and 3rd International Workshop, QASA 2014, Wroclaw, Poland, September 10–11, 2014. Data privacy management, autonomous spontaneous security, and security assurance. Springer International Publishing; 2015.
Garousi G, Garousi V, Moussavi M, Ruhe G, Smith B. Evaluating usage and quality of technical software documentation: an empirical study. 17th International Conference on Evaluation and Assessment in Software Engineering, Porto de Galinh, Brazil, 2013.
Garousi V, Felderer M, Mäntylä MV. EASE. In: Beecham S, Kitchenham B, MacDonell SG, editors. The need for multivocal literature reviews in software engineering: complementing systematic literature reviews with grey literature. ACM; 2016. 26:1–26:6.
Garousi, V., Felderer, M., Mäntylä, M. V., 2017. Guidelines for including the grey literature and conducting multivocal literature reviews in software engineering. abs/1707.02553
Gharibi, W., 2012a. Some recommended protection technologies for cyber crime based on social engineering techniques – phishing. CoRR abs/1201.0949
Gharibi, W., 2012b. Some recommended protection technologies for cyber crime based on social engineering techniques – phishing. CoRR abs/1201.0949
Ghosh S, Porras PA, Yegneswaran V, Nitz K, Das A. In: AAAI Workshops. ATOL: a framework for automated analysis and categorization of the darkweb ecosystem. AAAI Press; 2017.
Gordon S, Ford R. Cyberterrorism? Comput. Secur. 2002;21(7):636–47.
Green DM, Swets JA. Signal Detection Theory and Psychophysics. reprint. Los Altos, CA: Peninsula Publishers; 1989.
Griffith, V., Xu, Y., Ratti, C., 2017. Graph theoretic properties of the darkweb.CoRR abs/1704.07525
Gupta D, Rani R. Big data framework for zero-day malware detection. Cybern. Syst. 2018;49(2):103–21.
Hale, J., 2018. What is the dark web? From drugs and guns to the chloe ayling kidnapping, a look inside the encrypted network.
Han X, Wang L, Cui C, Ma J, Zhang S. Linking multiple online identities in criminal investigations: a spectral co-clustering framework. IEEE Trans. Inf. Forensics Secur. 2017;12(9):2242–55.
Harel D, Gery E. Executable object modeling with statecharts. IEEE Comput. 1997;30(7):31–42.
Haughey, H., Epiphaniou, G., Al-Khateeb, H. M., Dehghantanha, A., 2018. Adaptive traffic fingerprinting for darknet threat intelligence.CoRR abs/1808.01155
Hoofnagle CJ. In: Future of Privacy Forum Workshop on Big Data and Privacy: Making Ends Meet. How the fair credit reporting act regulates big data; 2013.
Huang K, Siegel M, Madnick S. Systematically understanding the cyber attack business: A survey. ACM Comput. Surv. 2018;51(4) 70:1–70:36.

Jasper SE. U.S. cyber threat intelligence sharing frameworks. Int. J. Intell. CounterIntelligence 2017;30(1):53–65.
Johnson RB, Onwuegbuzie AJ. Mixed methods research: a research paradigm whose time has come. Educ. Res. 2004;33(7):14–26.
Kandula S. Surviving DDoS attacks. Proceedings of the 2nd Symposium on NetworkedSystems Design and Implementation (NSDI05), 2005.
Kaspersky Lab daily, 2018. How to address incident response challenges.
Khan A. A novel approach to decoding: Exploiting anticipated attack information using genetic programming. Int. J. Knowl.-Based Intell. Eng. Syst. 2006;10(5):337–46.
Khelghati M. Deep web content monitoring. University of Twente, Enschede, Netherlands; 2016. Ph.D. thesis.. Base-search.net (ftunivtwente:oai:doc.utwente.nl:100466)
Kieseberg P, Segou OE, Roli F. Cyberroad: developing a roadmap for research in cybercrime and cyberterrorism. ERCIM News 2015;2015(102).
Kikuchi H, Fukuno N, Terada M, Doi N. Principal components of port-address matrices in port-scan analysis. In: Meersman R, Tari Z, editors. In: OTM Conferences (2). Springer; 2008. p. 956–68.
Kim, D., Kim, H. K., 2018. Automated dataset generation system for collaborative research of cyber threat intelligence analysis. CoRR abs/1811.10050
Kim K, Kim J, Hwang J. Ip traceback with sparsely-tagged fragment marking scheme under massively multiple attack paths. Cluster Comput. 2013;16(2):229–39.
Kitchenham B, Pearlbrereton O, Budgen D, Turner M, Bailey J, Linkman S. Systematic literature reviews in software engineering—A systematic literature review. Inf. Softw. Technol. 2008;51(1):7–15.
Kleinberg J. The small-world phenomenon: an algorithmic perspective. In: Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing. ACM; 2000. p. 163–70.
Kumar A, Mittra MK. Intrusion detection system an approach for finding attacks. Sci. Park Res. J. 2014;1(39).
Lakhotia A, Black P. Mining malware secrets. In: MALWARE. IEEE; 2017. p. 11–18.
Lauinger T, Pankakoski V, Balzarotti D, Kirda E. LEET. In: Bailey M, editor. Honeybot, your man in the middle for automated social engineering. USENIX Association; 2010.
Lauinger T, Pankakoski V, Balzarotti D, Kirda E. In: LEET. Honeybot, your man in the middle for automated social engineering; 2010.
League, H., 2018. What is surface web, deep web and dark web?
Leckie C, Ramamohanarao K. A probabilistic approach to detecting network scans. In: Stadler R, Ulema M, editors. In: NOMS. IEEE; 2002. p. 359–72.
Lewis KM, Hepburn P. Open card sorting and factor analysis: a usability case study. Electron. Lib. 2010;28(3):401–16.
Li Z, Sanghi M, Chavez B, Chen Y, Kao M-Y. Hamsa: fast signature generation for zero-day polymorphic worms with provable attack resilience; 2006.
Liu J, Fukuda K. An evaluation of darknet traffic taxonomy. JIP 2018;26:148–57.
Ma K, Sun R, Abraham A. Toward a lightweight framework for monitoring public clouds. In: Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on. IEEE; 2012. p. 361–5.
Ma M, Lin W, Pan D, Lin Y, Wang P, Zhou Y, Liang X. Data and decision intelligence for human-in-the-loop cyber-physical systems: reference model, recent progresses and challenges. Signal Process. Syst. 2018;90(8–9):1167–78.
Malwarebytes.com, 2019. All about malware.
Martine G, Rugg G. That site looks 88.46% familiar: quantifying similarity of web page design. Expert Syst. 2005;22(3):115–20.
Mataracioglu, T., Azkan, S., Hackney, R., 2015. Towards a security lifecycle model against social engineering attacks: Slm-sea.CoRR abs/1507.02458
Mazel J, Fontugne R, Fukuda K. TMA. In: Botta A, Sadre R, Bustamante F, editors. Identifying coordination of network scans using probed address structure. IFIP; 2016.
Meland PH, T øndel IA, Solhaug Bø. Mitigating risk with cyberinsurance. IEEE Secur. Privacy 2015;13(6):38–43.
Microsoft Philippines PR Team, 2017. Nus study: cybercriminals exploit pirated software to fuel malware infections in asia pacific.
Nabki MWA, Fidalgo E, Alegre E, de Paz I. Classifying illegal activities on Tor network based on web textual contents. In: Lapata M, Blunsom P, Koller A, editors. In: EACL (1). Association for Computational Linguistics; 2017. p. 35–43.
Narayanan, A., Shmatikov, V., 2009. De-anonymizing social networks.
Narita M, Kamada K, Ogura K, Bista BB, Takata T. A study of packet sampling methods for protecting sensors deployed on darknet. In: NBiS. IEEE Computer Society; 2016. p. 76–83.
Nassiri, A., 2018. Iot and DDoS: Cyberattacks on the rise.
Neu CV, Tatsch CG, Lunardi RC, Michelin RA, Orozco AMS, Zorzo AF. Lightweight IPS for port scan in OpenFlow SDN networks. In: NOMS. IEEE; 2018. p. 1–6.
Nishikaze H, Ozawa S, Kitazono J, Ban T, Nakazato J, Shimamura J. Large-scale monitoring for cyber attacks by using cluster information on darknet traffic features. In: Roy A, Angelov P, Alimi AM, Venayagamoorthy GK, Trafalis TB, editors. In: INNS Conference on Big Data. Elsevier; 2015. p. 175–82.
Nord RL, Ozkaya I, Schwartz EJ, Shull F, Kazman R. CSET @ USENIX Security Symposium. In: Eide E, Payer M, editors. Can knowledge of technical debt help identify software vulnerabilities?. USENIX Association; 2016.
Nunes E, Shakarian P, Simari GI. At-risk system identification via analysis of discussions on the darkweb. In: eCrime. IEEE; 2018. p. 1–12.
O'Riordan S, Feller J, Nagle T. A categorisation framework for a feature-level analysis of social network sites. J. Decis. Syst. 2016;25(3):244–62.
osint.it, 2015. OSINT, one important kind of intelligence.
Park H, Cho S, Kwon H-C. Cyber forensics ontology for cyber criminal investigation. In: Sorell M, editor. In: e-Forensics. Springer; 2009. p. 160–5.
Parveen P, McDaniel N, Weger ZR, Evans J, Thuraisingham BM, Hamlen KW, Khan L. Evolving insider threat detection stream mining perspective. Int. J. Artif. Intell. Tools 2013;22(5).
Pcmag Encyclopedia, 2019. What is internet domain?
Petersen K, Feldt R, Mujtaba S, Mattsson M. Systematic mapping studies in software engineering. In: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering. BCS Learning & Development Ltd.; 2008. p. 68–77.
Phillips P, Lee I. Mining top-k and bottom-k correlative crime patterns through graph representations. In: ISI. IEEE; 2009. p. 25–30.
Pokorny, Z., 2020. What are the phases of the threat intelligence lifecycle?
Porup, J., 2018. What is the Tor browser? How it works and how it can help you protect your identity online.
Rahul, Sujata. Host protection using process white-listing,deception and reputation services. IJIRIS 2018;V(Issue II):01–12.
Ring M, Landes D, Hotho A. Detection of slow port scans in flow-based network traffic. PLoS One 2018;13(9):1–18.
Ring M, Wunderlich S, Grdl D, Landes D, Hotho A. Flow-based benchmark data sets for intrusion detection. In: Proceedings of the 16th European Conference on Cyber Warfare and Security (ECCWS). ACPI; 2017. p. 361–9.

Rocha F. Insider threat: memory confidentiality and integrity in the cloud. Newcastle University, UK, British Library, EThOS; 2015. Ph.D. thesis.
Sanchez, J., Griffin, G., 2019. Whos afraid of the dark? Hype versus reality on the dark web.
Santos E, Nguyen H, Yu F, Kim KJ, Li D, Wilkinson JT, Olson A, Jacob R, Clark B. Intelligence analyses and the insider threat. IEEE Trans. Syst. Man Cybern. Part A 2012;42(2):331–47.
Schneider CM, Moreira AA, Andrade JS, Havlin S, Herrmann HJ. Mitigation of malicious attacks on networks. Proc. Natl. Acad. Sci. 2011;108(10):3838–41.
Scott-Hayward S, O'Callaghan G, Sezer S. SDN security: A survey. In: Future Networks and Services (SDN4FNS), 2013 IEEE SDN for; 2013. p. 1–7. doi:10.1109/SDN4FNS.2013.6702553.
Setola R, Luiijf E, Theocharidou M. Critical infrastructures, protection and resilience. In: Managing the Complexity of Critical Infrastructures. Springer, Cham; 2016. p. 1–18.
Shosha AF, Liu C-C, Gladyshev P, Matten M. Evasion-resistant malware signature based on profiling kernel data structure objects. In: Martinelli F, Lanet J-L, Fitzgerald WM, Foley SN, editors. In: CRiSIS. IEEE Computer Society; 2012. p. 1–8.
Sievert C, Shirley K. LDAvis: a method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Association for Computational Linguistics; 2014. p. 63–70.
Skopik F, Settanni G, Fiedler R. A problem shared is a problem halved: a survey on the dimensions of collective cyber defense through security information sharing. Comput. Secur. 2016;60:154–76.
Solano, P. C., Peinado, A. J. R.,. Socio-economic factors in cybercrime: statistical study of the relation between socio-economic factors and cybercrime.In: CyberSA. IEEE, pp. 1–4.
Soldani J, Tamburri DA, Heuvel W-JVD. The pains and gains of microservices: a systematic grey literature review. J. Syst. Softw. 2018;146:215–32.
Sorensen, S., 2012. Security implications of software-defined networks.
Stempfhuber M, Schaer P, Shen W. Enhancing visibility: Integrating grey literature in the sowiport information cycle. Ninth International Conference on Grey Literature: Grey Foundations in Information Landscape, 2008.
Symantec Employee, 2018. 10 cyber security facts and statistics for 2018.
Symantec employee, 2019. What is a man-in-the-middle attack?
Tang M, Li M, Zhang T. The impacts of organizational culture on information security culture: a case study. Inf. Technol. Manag. 2016;17(2):179–86.
Techopedia, 2019. Technology dictionary.
Techopedia, 2019. What does hacktivism mean?
Techopedia, 2019. What does honeypot mean?
Techopedia, 2019. What does spoofing mean?
Techtarget.com, 2019a. insider threat.
Techtarget.com, 2019b. watering hole attack.
Toch E, Bettini C, Shmueli E, Radaelli L, Lanzi A, Riboni D, Lepri B. The privacy implications of cyber security systems: atechnological survey. ACM Comput. Surv. 2018;51(2) 36:1–36:27.
Tounsi W, Rais H. A survey on technical threat intelligence in the age of sophisticated cyber attacks. Comput. Secur. 2018;72:212–33.
Veerasamy N, Grobler M. Logic tester for the classification of cyberterrorism attacks. IJCWT 2015;5(1):30–46.
Vidal C, Choo K-KR. Situational crime prevention and the mitigation of cloud computing threats. In: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 239; 2017. p. 218–33.
Wang M, Wang X, Shi J, Tan Q, Gao Y, Chen M, Jiang X. Who are in the darknet? Measurement and analysis of darknet person attributes. In: DSC. IEEE; 2018. p. 948–55.
Wisegeek, 2019. What is a web Crawler?
Wood G. The structure and vulnerability of a drug trafficking collaboration network. Soc. Netw. 2017;48:1–9.
Xu L, Zhan Z, Xu S, Ye K. Cross-layer detection of malicious websites. In: Proceedings of the Third ACM Conference on Data and Application Security and Privacy. ACM; 2013. p. 141–52.
Yang J, Shi G, Zheng Y, Wang Q. Data extraction from deep web pages. In: CIS. IEEE Computer Society; 2007. p. 237–41.
Yannikos Y, Schäfer A, Steinebach M. Monitoring product sales in darknet shops. Proceedings of the 13th International Conference on Availability, Reliability and Security. ACM, 2018.
Yegneswaran V, Barford P, Plonka D. On the design and use of internet sinks for network abuse monitoring. In: Jonsson E, Valdes A, Almgren M, editors. In: RAID. Springer; 2004. p. 146–65.
Zave P. In: Programming Methodology. An experiment in feature engineering. imported: Springer; 2003.

Giuseppe Cascavilla is a PostDoc researcher at the Eindhoven University of Technology, Jheronimus Academy of Data Science, in 's-Hertogenbosch, The Netherlands. Giuseppe completed his Ph.D. in Sapienza - University of Rome in 2018 with a thesis "Privacy Issues in Online Social Networks". His research interests lie mainly in cyber criminal activities monitoring in surface-deep-dark web, cyber threat intelligence, protection of cyber physical spaces, user profiling from social media activities, reidentification of personal emotions. Giuseppe is actually involved in the projects PRoTECT, ANITA, VISOR, and Cyber Physical Spaces Protection of Rotterdam Harbor where he is an active contributor and researcher.

Damian A. Tamburri is an Associate Professor at the Jheronimus Academy of Data Science, in 's-Hertogenbosch, The Netherlands. Damian completed his Ph.D. at VU University Amsterdam, The Netherlands in March 2014 one year in advance of his Ph.D. Contract. His research interests lie mainly in Complex Software Architectures (with a focus on Data-Intensive Architectures, Cloud & Microservices), Complex Software Architecture Properties (with a focus on Privacy & Security), and with Empirical Software Engineering (with a focus on Organizational, Social, and Societal aspects). Damian has been an active contributor and researcher in many EU FP7 and H2020 projects, such as S-Cube, MODAClouds, SeaClouds, DICE, RADON, SODALITE, ANITA, and more. The methodological, design, and infrastructure implementation experience with DevOps, infrastructure-as-code, TOSCA, and complex cloud styles featuring big data matured as part of the above projects and Damian's own research will play a key role in participation within COCLIA. In addition, he is an IEEE Software editorial board member, Voting member of the TOSCA TC as well as secretary of the IFIP TC2, TC6, and TC8 WG on Service-Oriented Computing".

Prof. dr. Willem-Jan van den Heuvel is a full professor in Information Systems and managing director of the European Research Institute of Services Science (ERISS). He is currently the scientific director of several NL and H2020 projects including the recently funded H2020 ANITA project focusing on evolutionary and collaborative software technology for digital and cyber crime-fighting. His research interests are at the cross-junction of software service systems and business process management with an emphasis on (global) networked enterprises. In particular, his expertise revolves around the following major research themes: business process management, Big data analytics, software service engineering (including service governance) and software legacy services modernization.