IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 19, 2019, accepted January 5, 2020, date of publication January 9, 2020, date of current version January 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2965257

The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends

JAVIER PASTOR-GALINDO, PANTALEONE NESPOLI, FÉLIX GÓMEZ MÁRMOL, AND GREGORIO MARTÍNEZ PÉREZ

Department of Information and Communications Engineering, University of Murcia, 30100 Murcia, Spain

Corresponding author: Javier Pastor-Galindo (javierpg@um.es)

This work was supported in part by the FPU Predoctoral Contract (FPU18/00304) granted by the Spanish Ministry of Science, Innovation and Universities, in part by the FPU Predoctoral Contract granted by the University of Murcia, in part by the Ramón y Cajal Research Contract (RYC-2015-18210) granted by the MINECO (Spain) and co-funded by the European Social Fund, and in part by the project SAFEMAN (*A unified management framework for cybersecurity and safety in the manufacturing industry*) with code *RTI2018-095855-B-I00*.

ABSTRACT The amount of data generated by the current interconnected world is immeasurable, and a large part of such data is publicly available, which means that it is accessible by any user, at any time, from anywhere in the Internet. In this respect, Open Source Intelligence (OSINT) is a type of intelligence that actually benefits from that open natureby collecting, processing and correlating points of the whole cyberspace to generate knowledge. In fact, recent advances in technology are causing OSINT to currently evolve at a dizzying rate, providing innovative data-driven and AI-powered applications for politics, economy or society, but also offering new lines of action against cyberthreats and cybercrime. The paper at hand describes the current state of OSINT and makes a comprehensive review of the paradigm, focusing on the services and techniques enhancing the cybersecurity field. On the one hand, we analyze the strong points of this methodology and propose numerous ways to apply it to cybersecurity. On the other hand, we cover the limitations when adopting it. Considering there is a lot left to explore in this ample field, we also enumerate some open challenges to be addressed in the future. Additionally, we study the role of OSINT in the public sphere of governments, which constitute an ideal landscape to exploit open data.

INDEX TERMS OSINT, cyberintelligence, cybersecurity, cyberdefence, challenges, national security, computer crime, computational intelligence, knowledge acquisition, social network services, software tools, data privacy, Internet.

I. INTRODUCTION

Open Source Intelligence (OSINT) consists in the collection, processing and correlation of public information from open data sources such as the mass media, social networks, forums and blogs, public government data, publications, or commercial data. Given some input data, together with the application of advanced collection and analysis techniques, OSINT continuously expands the knowledge about the target. In this way, the information found feeds the gathering process again to get closer to the final goal [1].

Nowadays, OSINT is widely adopted by governments and intelligence services to conduct their investigations and fight against cybercrime [2]. Nevertheless, it is not only utilised for state affairs, but rather applied to several different goals. Indeed, current research is focused on (but not limited to) three main applications which are represented in FIGURE 1 and are described next:

- *Social opinion and sentiment analysis:* Along with the boom of online social networks, it is possible to collect users interactions, messages, interests and preferences to extract non-explicit knowledge. The evidence accumulated from social media is far-reaching and widely advantageous [3]. Such collection and analysis could be applied, for instance, to marketing, political campaignsor disaster management [4].
- *Cybercrime and organized crime:* The open data is continuously analyzed and matched by OSINT processes in order to spot criminal intentions at an early stage. Taking into account adversaries' patterns and relationships between felonies, OSINT is able to provide security forces with an opportunity to promptly detect

The associate editor coordinating the review of this manuscript and approving it for publication was Luis Javier Garcia Villalba.

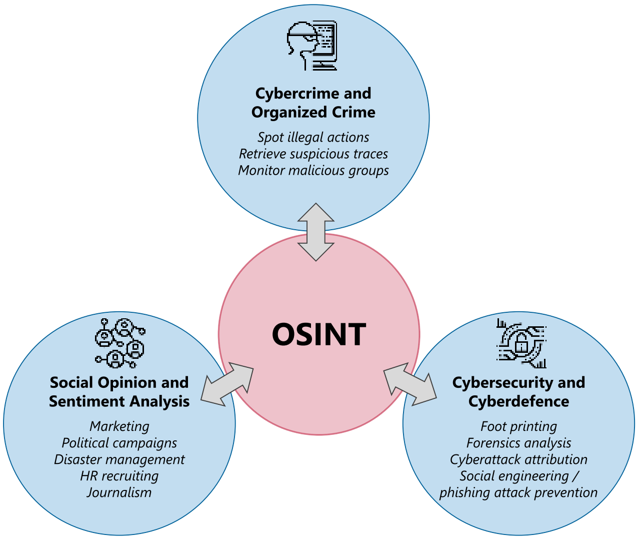

FIGURE 1. **OSINT principal use cases.**

illegal actions [5]. In this direction, by exploiting the open data, it would be possible to track the activity of terrorist organizations, which are increasingly active on the Internet [6], [7].

- *Cybersecurity and cyberdefence:* ICT (*Information and Communication Technology*) systems are continuously attacked by criminals aiming at disrupting the availability of the provided services [8]. Research becomes hence crucial to defend those systems from cyberattackers, concretely by facing the challenges that are still open in the field of cybersecurity [9]. In this sense, data sciences are not only being applied to the footprinting in pentestings, but also to the preventive protection of organizations and companies. Concretely, data mining techniques may help by performing analysis of daily attacks, correlating them and supporting decision making processes for an effective defense, but also for a prompt reaction [10]. In the same way, OSINT can be also considered in this context as a source of information for tracebacks and investigations. Forensic digital analysis [11] can incorporate OSINT to complement the digital evidences left by an incident.

In addition to those, OSINT can be applied to other contexts. In particular, one may extract relevant information by performing social engineering attacks. Ill-motivated entities leverage publicly-available information released online (e.g., on social networks) in order to create appealing hooks to capture the target [12]. Moreover, it is possible to perform automatic veracity assessment on the open data aiming at disclosing fake news and deepfakes, among others [13].

Nonetheless, it is important to notice that the utilization of public data has also compromising issues. On the one hand, the EU General Data Protection Regulation (GPDR) limitates the processing of personal data related to individuals in the EU zone [14]. On the other hand, there is a strong ethical component which is linked to the users' privacy. In particular, the profiling of people [15] could reveal personal details such as their political preference, sexual orientation or religious beliefs, amongst others. Additionally, the exploitation of such vast amount of information may lead to abuse, resulting in harming innocents through cyberbullying, cybergossip or cyberaggressions [16].

The paper at hand, which is an extension of the work proposed in [17], encompasses the present and future of OSINT by analyzing its positive and negative points, describing ways of applying this type of intelligence, and enunciating future directions for the evolution of this paradigm. In addition, a more detailed description of different techniques, tools and open challenges is presented in this work. Furthermore, we propose the integration of OSINT within the DML (*Detection Maturity Level*) model to address the attribution problem from a different perspective in the context of cyberattacks investigations. We also introduce sample workflows to facilitate the understanding and use of OSINT to gather valuable information starting from basic inputs.

In addition, our purpose is to stimulate researches and advances in the OSINT ecosystem. The scope of such ecosystem is quite wide, spanning from psychology, social science to counterintelligence and marketing. As we have seen so far, OSINT is a promising mechanism that concretely improves the traditional cyberintelligence, cyberdefence and digital forensic fields [18]. The impact that this methodology could have on society thanks to current technology and the large number of open sources is still unexploited. There is still a long way ahead to explore in this topic, and this article presents some future appealing research lines.

The remainder of this paper is organized as follows. SECTION II offers a review of recent research works in the field of OSINT. SECTION III discusses the motivation, pros and cons of the development of OSINT. SECTION IV explains the principal OSINT steps and practical workflows to carry them out. Then, SECTION V includes an in-depth description of OSINT-based collection techniques and services. SECTION VI analyzes and compares some OSINT tools that automatize the OSINT collection and analysis of information. SECTION VII proposes the integration of OSINT in the investigation of cyberattacks. SECTION VIII focuses on the impact of OSINT within a nation, not only for the sake of its internal cyberdefence operations, but also as a beneficiary of transparency policies. Spain is specifically taken as a reference for affinity and contextualized with the rest of the world. SECTION IX poses some open challenges regarding research in OSINT. Finally, SECTION X concludes with some key remarks, as well as future research directions.

II. STATE OF THE ART

In recent years, with the advances of big data and data mining techniques, the research community has noticed that open data represents a powerful source of analyzing social behaviors and obtaining relevant information [19]. Next we describe some remarkable works pivoting around each of the three aforementioned principal use cases for OSINT.

With regards to the use of OSINT for **extracting social opinion and emotions**, Santarcangelo *et al.* [20] proposed a model for determining user opinions about a given keyword through social networks, specifically studying the adjectives, intensifiers and negations used in tweets. Unfortunately, it is a simple keyword-based solution only designed for Italian language, not taking into account semantic issues. On the other hand, Kandias *et al.* [21] could relate people usage of social networks (in particular, Facebook) to their stress level. However, the experiments were carried out only with 405 users, while nowadays there is a chance of processing much larger amounts of data. Another interesting study is conducted in [22], where authors applied Natural Language Processing (NLP) to WhatsApp messages in order to possibly prevent the occurrence of mass violence in South Africa. Unfortunately, the investigation is limited to text messages, thus excluding vital information which can be disclosed through multimedia material.

In the context of **cybercrime and organized crime**, there are several works that explore the application of OSINT for criminal investigations [23]. For example, OSINT could increase the accuracy of prosecutions and arrests of culprits with frameworks like the one proposed by Quick and Choo [11]. Concretely, authors apply OSINT to digital forensic data of a variety of devices to enhance the criminal intelligence analysis. In this field, another opportunity that OSINT yields is the detection of illegal actions as well as the prevention of future crimes such as terrorist attacks, murders or rapes. In fact, the European projects ePOOLICE [24] and CAPER [25] were designed to develop effective models for scanning open data automatically in order to analyze the society and detect emerging organized crime. In contrast to the previous mentioned projects, whose proposals were not practically used in real cases, Delavallade *et al.* [26] describe a model based on social networks data that is able to extract future crime indicators. Such model is then applied to the copper theft and to the jihadist propaganda use cases.

From the point of view of **cybersecurity and cyberdefence**, OSINT represents a valuable tool for improving our protection mechanisms against cyberattacks. Hernández *et al.* [27] propose the use of OSINT in the Colombian context to prevent attacks and to allow strategic anticipation. It includes not only plugins for collecting information, but also machine learning models to perform sentiment analysis. Moreover, the DiSIEM european project [28] maintains as a first goal the integration of diverse OSINT data sources in current SIEM (*Security Information and Event Management*) systems to help reacting to recently-discovered vulnerabilities in the infrastructure or even predicting possible emerging threats. In addition, Lee and Shon [29] also designed an OSINT-based framework to inspect cybersecurity threats of critical infrastructure networks. However, all these approaches have not been applied to real world scenarios, thus their effectiveness remains questionable.

Extending the dissertation to other application fields, in [30] authors demonstrate how to passively recollect significant information on organizational employees in an automated fashion. Such information is then related to the analysis of the so-called *social engineering attack surface*, showing the effective feasibility of the proposed approach. Then, the authors propose a set of potential countermeasures, including a publicly-available social engineering vulnerability scanner which companies may leverage in order to reduce the exposure of their employees.

Furthermore, a systematic review of approaches, methodologies and tools which are proposed by the academy to conduct automatic veracity assessment of publicly-available data is performed in [31]. Specifically, the authors studied 107 research items between 2013 and 2017 to argue on the state-of-the-art of veracity assessment, which has become a great concern during the last decade due to the spread of fake news and deepfakes. In this direction, the authors outline the relative immaturity of this field, identifying several challenges which will characterize future research trends.

III. OSINT ADVANTAGES AND SHORTCOMINGS

The fields of application of OSINT are numerous and the solutions being developed under this paradigm are increasing. However, behind this methodology there is a trade-off that developers and engineers have to deal with. From a technical point of view, as we can see in TABLE 1, OSINT exposes a number of benefits, but it has to deal with some restrictions too, which are detailed next.

A. OSINT BENEFITS

1) HUGE AMOUNT OF AVAILABLE INFORMATION

There is currently a large volume of worthwhile open source data to be analyzed, correlated and linked [32]. This includes social networks, public government documents and reports, online multimedia content, newspapers and even the Deep web and the Dark web [33], among others. Actually, both the Deep Web and the Dark Web (the latter circumscribed within the former) contain even more information than the Surface Web (i.e., the Internet known by most users) [34]. In order to be able to access these networks, it is necessary to use specific tools since their contents are not indexed by traditional search engines.

Unlike the Surface Web and most of the Deep Web, the Dark Web offers anonymity and privacy to users who utilize it. This property facilitates criminals to employ this network to surf, conduct their searches and publish with illegitimate purposes while hiding their identity. Therefore, the Dark Web is an ideal source to apply OSINT and fight against cybercrime, organized crime or cyberthreats. On the other hand, the pursuit and de-anonymization of these people are current non trivial challenges for OSINT to properly work [35].

2) HIGH COMPUTING CAPACITY

Advances in computer architecture, processors and GPUs (graphic processing units) enable to carry out labor-intensive

TABLE 1. OSINT pros and cons in a nutshell.

Pros ✓	Cons ✗
Huge amount of available information	Complexity of data management
High capacity of computing	Unstructured information
Big data and machine learning	Misinformation
Complementary types of data	Data sources reliability
Flexible purpose and wide scope	Strong ethical/legal considerations

operations in terms of collection, processing, analysis and storage [36]. Thanks to this feature, we have the opportunity to apply OSINT considering large amounts of public information and mixing a high number of data sets, relationships and patterns from different types of open sources, while applying advanced processing and analysis techniques.

3) BIG DATA AND MACHINE LEARNING

Emerging proliferation of data analysis and data mining techniques, as well as machine learning algorithms, which can automate and make investigation and decision making processes more intelligent and efficient [36]. It allows spotting complex correlations that are naturally unpredictable to humans. This point will be key in future OSINT activities, as it will mark the difference between human-driven and artificial intelligence-led research. By incorporating those techniques, the process of collection and analysis will definitively improve, thus resulting in accurate investigations close to our goal. Additionally, government counterintelligence agencies can leverage such paradigm to further enhance the quality of managed information and, consequently, the battle against terrorist organizations [37].

4) COMPLEMENTARY TYPES OF DATA

Possibility of feeding OSINT with other types of information [38]. The inherent structure of the system is open enough to include data that has not actually been obtained from open sources. This fact means that OSINT can be even more effective if we are able to add external pieces of information to complement investigations. For example, Law Enforcement Agencies could take advantage of citizens collaboration to feed OSINT searches, intelligence services could leverage classified information about cybercriminals or incidents to enrich OSINT investigations, or even common users could combine OSINT with social engineering to profile their target.

5) FLEXIBLE PURPOSE AND WIDE SCOPE

Due to the nature of OSINT, investigations can be extended to lots of problems and can collect pieces of information all over the cyberspace. This paradigm could be used for economic, psychological, strategic, journalistic, labor or security aspects, among others. In particular, we could highlight the benefits in the field of crime and cybersecurity, where OSINT could monitor suspicious people or dangerous groups, detect influencing profiles related to radicalization, study worrying trends of the society, support the attribution of cyberattacks and crimes, enhance digital forensic analysis, etc. [5], [18].

B. OSINT LIMITATIONS

1) COMPLEXITY OF DATA MANAGEMENT

The quantity of data is huge and, consequently, it is challenging to handle it efficiently and effectively [39]. It is beneficial for OSINT to consider as much information as possible, but also to have advanced techniques and significant resources to ensure high quality collection, processing and analysis.

2) UNSTRUCTURED INFORMATION

The public information available on the Internet is inherently massively disorganized. This means that the data collected by OSINT is so heterogeneous that turns it tough to classify, link and examine such data in order to extract relevant relationships and knowledge [4]. In this sense, OSINT requires mechanisms such as data mining, Natural Language Processing (NLP), or text analytics to homogenize the unstructured information in order to be able to exploit it.

3) MISINFORMATION

Social networks and communication media are flooded with subjective opinions, *fake news* and canards [4]. For this reason, the existence of inaccurate information has to be taken into account in the implementation of OSINT mechanisms and should not drive the propagation of the search. OSINT activities should always deal with reliable information and follow trusted exploration lines to ensure positive and convincing outcomes [40].

4) DATA SOURCES RELIABILITY

The trustworthiness and authority of the information are indeed the key for successful OSINT investigations [41]. Ideally, the collected data should come from authoritative, reviewed and trusted sources (official documents, scientific reports, reliable communication media) [39]. In practice, OSINT will also coexist with subjective or non-authoritative sources, such as the content of social networks or manipulated media [42]. Even though this type of sources is more prone to misinformation, it is actually where more knowledge can be extracted to investigate people, groups or companies. If the credibility of the open sources of information represents indeed a limitation, it becomes even more challenging considering the possible ambiguity of users' queries to retrieve the desired information [43].

5) STRONG ETHICAL/LEGAL CONSIDERATIONS

Numerous concerns about privacy, respect and personal integrity emerge with the development of OSINT [44]. In this direction, it has to be noted that the question of whether OSINT constitutes an ethical issue is generally situated within the area of the ethics of intelligence collection [45]. On the one hand, although publicly accessible, OSINT has the power to disclose information that is not explicitly posted on the web. Uncovered results should respect users' privacy and not reveal intimate and personal issues [15], while taking into account current related regulations (such as GPDR [14]). To this extent, aspects such as sexual orientation, religious beliefs, political inclination or compromising behaviours can be inferred from the Internet, and this disclosure process can be problematic in many countries today. On the other hand, the scope of OSINT-based searches should be, by definition, limited to open data sources. Under no circumstances access controls or authentication methods can be bypassed to extract knowledge.

IV. OSINT WORKFLOWS

OSINT, like any other type of intelligence, has a well-defined and precise methodology. From our scientific-technical point of view, we are particularly interested in three steps.

Firstly, in the **collection** phase, publicly available data is retrieved from relevant open sources according to the target or objective. In particular, the Internet is the resource par excellence due to the volume of existing material and easy accessibility. The collection process is particularly relevant because from this stage onwards the whole process of intelligence generation is triggered.

Then, in the **analysis** phase, the collected raw material is treated to generate valuable and comprehensible information. The data by itself is not useful, so it has to be interpreted to obtain the first facts derived from an in-depth analysis.

Finally, in the **knowledge extraction** process, the information purified previously is taken as input for more sophisticated inference algorithms. Thanks to the computational advances of current era, it is possible to detect patterns, profile behaviours, predict values or correlate events.

It is worth mentioning that the second and third steps comprise technologies widely used and known in the context of data mining. However, the OSINT collection approach differs from current data-driven services. Nowadays, common data analysis applications gather as much information as possible from pre-defined data sources and implement clear gathering processes. On the contrary, OSINT solutions should collect specific facts from the sea of all possible and reachable open resources.

In order to face this latter challenging uncertainty and go one step further, we propose in FIGURE 2 a practical framework to carry out OSINT-based investigations. We have included those exploration paths which are worthwhile to follow for optimizing the analysis of collection results and maximizing the extraction of knowledge. This high abstraction scheme includes the most clear transactions, representative elements and outstanding operations.

A. OSINT COLLECTION

Before the analysis and intelligence extraction steps, the investigator has to expand the dataset about the target. With this aim, we propose some OSINT techniques to represent different collection strategies. In particular, we have considered *search engines, social networks, email address, username, real name, location, IP address* and *domain name* OSINT techniques (as we will further describe in SECTION V). Under each one, there will be innumerable OSINT services with similar ways of collecting data.

In this phase, it is assumed that, at least, an atomic piece of data about the target is available (e.g., real name, username, email address, etc.). From that initial seed and according to its nature, the investigator applies the most suitable OSINT techniques to derive more data. In this sense, the results obtained with a specific technique are a *data transfer* to be used by another type of technique. These represented transactions illustrate possible ways of propagating the investigation, where the output of the technique of origin becomes the input to feed the technique of destination.

B. OSINT ANALYSIS

The continuous iterations through the different OSINT techniques should be analyzed and understood to generate valuable information. There is an increasing amount of analysis techniques in the literature to do this task [46], highlighting below those appealing procedures which are applicable in our scenario:

- *Lexical analysis*: Raw data should be examined to extract entities and relations from text. It is essential to apply translation processes to the language used in the OSINT investigation [47] and filter noise which does not add value from sentences that do not add value.
- *Semantic analysis*: Having a bag of words is not useful if the meaning is not extracted [48]. With this purpose of understanding data, natural language processing algorithms are being used nowadays [49]. In addition, sentiment analysis techniques permit the contextualization of subjective posts or opinions to classify the emotional status of the author (e.g, positive, negative or neutral). Finally, truth discovery procedures address the challenging task of resolving conflicts in multi-source data which stands opposing positions on the same subject [50].
- *Geospatial analysis*: Recollected data from social networks, events, sensors or IP addresses are worthwhile to be analyzed from a location-based perspective. In this sense, the usage of maps or graphs facilitates the representation and comprehension of data [51], as well as extracting meaningful connections between incidents or persons.
- *Social media analysis*: The features brought by modern social media allow researchers to carry out in-depth analysis of users [52]. In such a scenario, the analysis of

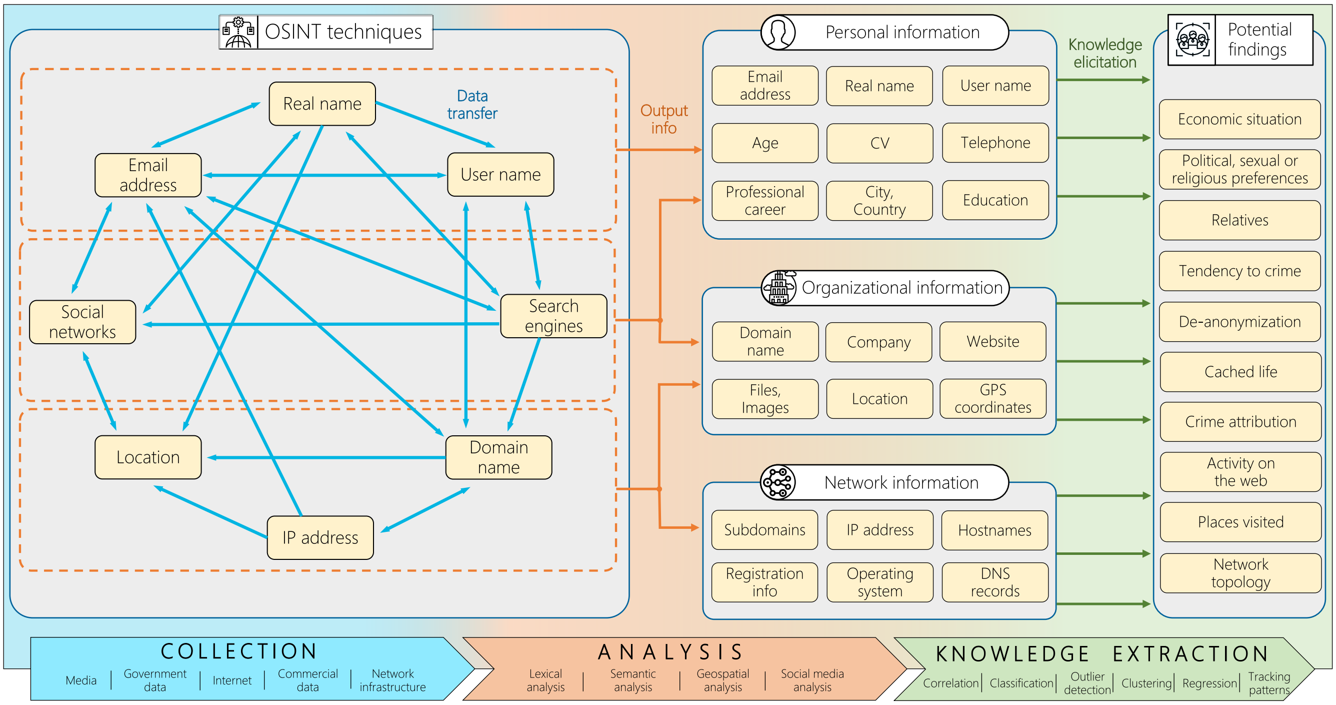

FIGURE 2. **Principal OSINT workflows and derived intelligence.**

social data allows the creation of a network of contacts, interactions, places, behaviours and tastes around the subject.

The results of launching the aforementioned techniques are considered as *output info* and are categorized into three main groups:

- The *personal information* fuses the person identity details which are mainly obtained from the real name, email address, user name, social networks and search engines techniques.
- The *organizational information* is formed by aspects of a team or company composed of individuals. It is essentially collected by means of social networks, search engines, location, domain name and IP address techniques.
- The *network information* covers technical data of systems and communication topologies which is usually achieved through location, domain name and IP address techniques.

Logically, these three blocks of information can be expanded with more elements. Moreover, a single investigation may have different types of *output info* that complement each other.

C. OSINT KNOWLEDGE EXTRACTION

The value of the information collected so far is unquestionable. However, the intelligence extraction of those findings leads actually to what will provide an attractive recognition of the target [53]. To this end, we consider the *knowledge elicitation* as the treatment of the analysis results (*output info*) making use of data mining and artificial intelligence techniques. In the following we mention some really promising technologies at this stage:

- *Correlation*: Detection of relationships between people, events or pieces of data in general [54]. Strong related features are specially valuable to reveal those non-explicit associations existing in the dataset.
- *Classification*: The data can be divided in groups according to predefined categories (supervised learning) [55]. This technique permits the organization of large amounts of information for more effective knowledge extraction [56].
- *Outlier detection*: This procedure analyzes the dataset and detects anomalies in it [57]. They are particularly interesting for the observation of malignant agents, whose behaviour or actions differ from the general population.
- *Clustering*: It assigns pieces of data into clusters, being able to consider big amount of conditions or heuristics [58]. This could reveal, for example, different ways of behaving in the network, various types of online profiles or categorizing forms of attacking individuals, organizations or infrastructures [59] without knowing the existence of that diversity beforehand (unsupervised learning).
- *Regression*: The main objective of this technique is to forecast or predict numeric values or facts [60]. For example, a linear regression returns a value attending to a linear function, a neural network is a structure that maps complex combinations of inputs to an output, or deep learning that is made up of several layers that combine and make operations with the input.

- *Tracking patterns*: Differing from anomaly detection, pattern recognition is a process for detecting regularities in data [61]. The methods mentioned above can be included in this knowledge-discovery broad concept. In fact, any artificial intelligence technique is suitable for open data knowledge extraction.

These intelligent techniques allow inferring abstract, complex and juicy issues about the target that are not explicitly published on the Internet [62]. However, this process poses several challenges, mainly residing in researching and developing this knowledge extraction process to identify, profile or monitor criminals, recognize and explore malicious organizations or uncover and attribute cybernetic incidents. In addition, several privacy considerations arise due to the powerful inferences that are potentially achievable. The extracted knowledge about a person, company or organizations may be specially sensible and its manipulation indirectly leads to ethical and legal problems (specifically addressed in SUBSECTION IX-F). Indeed, we should never lose sight of the fact that these techniques could be even misused to directly harm people or groups (deeper analysis in SUBSECTION IX-G).

V. OSINT COLLECTION TECHNIQUES AND SERVICES

As it has been shown, OSINT is quite promising and powerful, but its implementation is also challenging. . In fact, the first consideration is that it precises data as departure point. Fortunately, the volume of raw data is not a problem nowadays due to the existence of the Internet. In addition, there is also an increasing number of applications, known in this context as OSINT services, that precisely facilitate the gathering on the web.

In the following, a summary of the most common *OSINT techniques is presented*. Within each technique, the most outstanding associated *OSINT services* at the time of writing are shown, giving hints on how to effectively exploit their potentialities. It is worth mentioning that OSINT services are ephemeral and can even increase or decrease. On the contrary, the OSINT technique is a broader concept that will endure over time.

A. SEARCH ENGINES

Google, *Bing* or *Yahoo* search engines, among others, are well known and widely used tools. The traditional use of them is the simplest way of applying OSINT. These engines search within the World Wide Web given a textual query trying to provide information that matches with the input, working really well and returning valuable information to the user.

Nevertheless, the number of results can be so overwhelming that it can even be counterproductive for the user. For that reason, a good investigator should know how to specify the requests within a search engine according to the desired outcome. Services like *Google* or *Bing* support filters to refine searches,[1] and retrieve exactly the type of information we are interested in. For instance, the use of *“”* permits exact-matches, *OR* and *AND* act as logical operators, or * as a wildcard. It also allows the introduction of conditions like *filetype* to specify a certain file type, *site* to limit results to those from a specific website, or *intitle* to find pages with certain keywords within their title. TABLE 2 contains some operators that can be used to refine *Google* and *Bing* searches.

Yahoo, in turn, does not permit specific filters, but we can restrict the date, language or country of the results. The case of the *DuckDuckGo* search engine is specially interesting because it does not track the user, nor it targets the IP address or the search history. This privacy-preserving approach makes the findings homogeneous for all users, regardless of habits, preferences, location, or search history.

Moreover, some search engines have been designed for specific territories. *Yandex* is well-known in Russia and Eastern Europe, and implements search operators[2] to restrict the search by URL, file type, language, date, and so on. *Baidu* is another specific search service widely used in Asia. It includes not only the typical keyword search bar, but additional worthy resources for OSINT such as a social network, a section of questions and answers, a virtual library or an encyclopedia, among others. There are also search engines for the Arabic community such as *Yamli* or *Eiktub*, but they are much less employed. This type of services is particularly interesting in investigations about people, groups and companies belonging to specific communities.

Finally, it is mandatory to know specific search engines to browse the Dark Web. OSINT investigations against drug traffic, child pornography, weapon sales or terrorism are very benefited from exploring these not-so-popular resources. To this end, *Ahmia* and *Torch* are search engines available for use within the Tor anonymous network [63]. However, the researcher will have to deal with the anonymity of this network and sites.

B. SOCIAL NETWORKS

Nowadays, the exposure of the daily life of individuals and organizations in social networks is evident. Any curious person has realized that lot of personal information can be found with no advanced knowledge needed about these platforms. As shown in TABLE 3, these applications offer precise search possibilities in the context of OSINT. Next we describe some of the most known and used social networks worldwide.

Facebook is a social network spread all over the world with millions of users. It could be considered a diary of society, where one can find very valuable personal information for OSINT investigations. The profile of our target can reveal his employment, education, age, location, visited places or liked groups, among others. The photos and publications may also help us contextualize the company or person we are investigating, the areas it frequents or the type of activities he/she realizes. In addition, it is also possible to search by

[1] https://support.google.com/websearch/answer/2466433

[2] https://yandex.com/support/search/query-language/search-operators.html

TABLE 2. Some Google/Bing filters for advanced search.

Google/Bing filter	Search operator	Example of use
Force an exact-match search	" "	`"University of Murcia"`
Exclude a term or phrase	-	`university murcia -catholic`
Search for X or Y	`OR, |`	`university murcia|cartagena`
Search for X and Y (used by default)	`AND`	`university AND of AND murcia`
Use of a wildcard	`*`	`university of *`
Search for a range of numbers	`..`	`university murcia 2010..2019`
Group terms or search operators	`()`	`"university of (murcia|cartagena)"`
Search within a given domain	`site:`	`university murcia site:um.es`
Search for a certain file type	`filetype:`	`university murcia filetype:pdf`
Search in page titles	`intitle:`	`university intitle:umu`
Search in URLs	`inurl:`	`university inurl:um`
Search in the text of the pages	`intext:`	`university intext:murcia`
Search the most recent cached version of a page	`cache:`	`cache:um.es`

TABLE 3. Potential of various social networks.

Social Network	Type	Scope	Main potential for OSINT
4chan	Online community	Worldwide	Users interested in illicit activities
Badoo	Dating	Worldwide	Intimate and personal details
Cloob	Social connections	Iran	Personal profile, posting and community membership
Draugiem	Social connections	Latvia	Personal profile, publications in blogs, group membership
Facebook	Social connections	Worldwide	Personal profile, preferences and places visited
Facenama	Social connections	Iran	Personal profile, publications, photos and videos
Flickr	Photo-sharing	Worldwide	Activities, hobbies, places and personal relationships
Instagram	Social connections	Worldwide	Habits, locations and personal relationships
LinkedIn	Business	Worldwide	Professional profile, education, skills and languages
Mixi	Social connections	Japan	Personal profile, interests and opinions
Odnoklassniki	Social connections	Mainly Russia	Personal profile of adults, past and present friendships
Qzone	Social connections	Mainly China	Personal profile, preferences, habits
Reddit	Online community	Worldwide	Users trends, behaviors, and publications
Renren	Social connections	Mainly China	Personal profile of students, friendships and discussions
Taringa!	Social connections	Mainly Latin America	Personal profile, publications and community membership
Tinder	Dating	Worldwide	Intimate and personal details
Tumblr	Photo-sharing	Worldwide	Activities, hobbies, places and personal relationships
Twitter	Social connections	Worldwide	Personal profile, opinions and publications
VKontakte (VK)	Social connections	Mainly Russia	Personal profile, preferences and publications
Weibo	Social connections	Mainly China	Personal profile, opinions and publications
YouTube	Video-sharing	Worldwide	Video content, opinions and comments of subscribers

location when the real name is not known, being able to ultimately find the profile of our target.

YouTube is a video-based platform where big communities are conformed around shared interests. It is not only valuable the content uploaded by an specific user (themes, images, scenes, places, and people appearing in videos), but also the opinions and comments of subscribers.

Twitter is mainly utilized for live communication where it is common to find personal publications through an ordered timeline. Apart from the personal information revealed by the profile, it is particularly interesting the extraction of the opinions from published *tweets*, the relationships with followed and follower users or the *likes* in certain publications. From this type of interactions, an OSINT investigator can infer the orientation of the target on certain issues, the interests and preferences of an organization, or how dangerous a person might be. Additionally, a user-friendly interface[3] is available where it is possible to search on the whole platform by keywords, exact phrases, hashtags, language, date and so on. Thus, we can even define explorations through users, mentions or responses.

Instagram is also widespread in the modern society as a mean of sharing photos. The places, persons and activities shown in pictures can also assist us in profiling our target. The location is a quite sensitive data that is frequently shared on this platform. In this sense, we can also

[3] twitter.com/search-advanced

mention more specific photo-sharing services like *Tumblr* or *Flickr*.

LinkedIn is the most popular site in the context of business-related social networking. It permits searching by real name, company, organization, title or location. In this case, the professional profiles can reveal full contact data, including email addresses and cellular telephone numbers. In addition, we can also extract information about the employment, education, skills, languages and business relationships.

It is also worth considering those dating websites used to contact people in search of a mate. Unlike other social networks, where many users restrict their personal details, more intimate aspects are usually revealed in here. For this reason, services like *Tinder* or *Badoo* are useful for investigating the background information, personal character, interests, preferences or behaviour of the target.

Finally, it is possible to browse online communities which are very similar to social networks. The posts and topics of these forums generate interesting interactions to be analyzed by OSINT [64]. *Reddit* or *4chan* are big communities which host countless threads of discussion and opinion where really personal and private information about the target can be identified. However, in these websites users are commonly anonymous. Additionally, it is not rare to find illicit content of bullying, pornography or threats.

On the other hand, there are also some social networks which are typically used within specific regions. The following services are specially important in some countries.

Qzone, *Weibo* and *Renren* are some of the most used social networks in China. The first one is a very customizable platform where users publish blogs, diaries, photos or music which reveal details about the person. The second one has similar features to *Twitter*, but also including polls, file sharing and stories (temporal photo and video sharing). The last one is widespread among college students. Those OSINT investigations whose target is a Chinese person can get a valuable profit from these sites.

There are also social networks to interconnect Russian compatriots and eastern European citizens. In this regard, *VKontakte*, also known as *VK*, is very popular. The functionalities, and even the appearance, are quite similar to *Facebook*. Users are able to stay involved with friends, participate in online communities, post messages, photos, and videos in private or public pages, and even share files. Another Russian site to highlight is *Odnoklassniki*, mainly used by adults. In fact, the main purpose of its users is to have an online profile, keep in touch with real-life friendships and search former companions or past friends. In this sense, OSINT can be conducted to discover people-to-people connections from the past to now.

In Japan, *Mixi* is a very common social networking site in society. Apart from typical functionalities, we could highlight the possibility to make reviews to products, create personal blogs within the platform, participate in communities or manage music preferences and listening habits.

For Spanish-speaking countries, specially Latin America, *Taringa!* is a well known social platform for sharing photos, videos and news with friends. In addition, users are able to create communities, play online games or share music.

Finally, due to the existing censorship with external services, in Iran the most popular local social networks are *Facenama* and *Cloob*. The first is mainly used for sharing posts, photos and videos whereas the second includes community discussions, photo sharing, posting or chat rooms. Something similar about censorship occurs in Latvia, where *Draugiem* is widely used to share contents and communicate online.

C. EMAIL ADDRESS TECHNIQUE

Searching by a person's real name can be frustrating due to potentially duplicated names, so it is sometimes worth starting from an email address which is unique and achieves much better results at a faster pace. There are some interesting OSINT services, as it is shown in TABLE 4, that work with an email address as an input.

First of all, *Hunter* can be used to determine whether an email address is valid or not. Then, *Have I Been Pwned* informs whether a given email address is contained in public breaches (so that it has been compromised at some point). In particular, it is worth mentioning that the investigator can browse the list of sites where the email address was compromised. These services are potential sources for finding public information about the owner. Another worthwhile page is *Pipl*, which works really well to find information about the owner of an email address such as the real name, usernames, address, telephone number, education, professional career, etc.

D. USERNAME TECHNIQUE

The nicknames used for online services are also a good way to collect information regarding a person, as shown in TABLE 5. Visiting these services will allow an investigator to automatically check a username in several websites at the same time to identify more sources of information.

The services *KnowEm*, *Name Chk*, *Name Checkr*, or *User Search* verify the presence of a given username on the most popular social networks and domains.

NameVine, in turn, provides an interesting feature that helps when trying to guess an exact username. Concretely, it suggests profiles for the top ten social networks which partially match with the given username. This real time solution offers a fast verification of username variants (for instance, changing the final number of the nickname) instead of launching time-consuming queries repeatedly with other services.

The website *Lullar* uses a different approach. It automatically generates URLs to visit the username profile in different social networks without checking if they exist. If a link works, then the profile exists for that social network, whereas if it is broken it obviously means the opposite. In addition to speeding up manual checking, the most useful application would be to explore possible usernames when the one we have is

TABLE 4. Utility of the OSINT services belonging to the email address technique.

Email address OSINT service	URL	Main output
Hunter	`hunter.io`	Validity and availability
Have I Been Pwned	`haveibeenpwned.com`	Appearance in public data breaches
Pipl	`pipl.com`	Personal information about the owner

TABLE 5. Utility of the OSINT services belonging to the username technique.

Username OSINT service	URL	Main output
KnowEm	`knowem.com`	Presence in social networks, domains and online communities
Name Chk	`namechk.com`	
Name Checkr	`namecheckr.com`	
User Search	`usersearch.org`	
NameVine	`namevine.com`	Suggestions of alternative similar usernames
Lullar	`com.lullar.com`	Availability in social networks

TABLE 6. Utility of the OSINT services belonging to the real name technique.

Real name OSINT service	URL	Main output
Pipl	`pipl.com`	Personal information
That's Them	`thatsthem.com`	Personal details, education, professional career, skills, locations, and relatives.
Spokeo	`spokeo.com`	
Fast People Search	`fastpeoplesearch.com`	
Nuwber	`nuwber.com`	
Cubib	`cubib.com`	
Peek You	`peekyou.com`	
Yasni	`yasni.com`	Social networks profiles
Family Search	`familysearch.org`	Kinship information, relatives
GENi	`geni.com`	
Family Tree Now	`familytreenow.com`	
True People Search	`truepeoplesearch.com`	

questionable or partial. When the initial URL fails, similar or alternative users are often listed by the social networks which can be used to identify the entire existing username.

E. REAL NAME TECHNIQUE

Searching a target real name could also yield good results, as shown in TABLE 6. Apart from social networks, particular services are capable of revealing home addresses, telephone numbers, email accounts, usernames, among others.

We could highlight *Pipl* as the website that returns more information given a first and last name. Due to possible multiple results for the same real name, it is possible to refine the search by including additional aspects of the person such as email, phone, country, state, city, username or age.

That's Them also offers a remarkable output containing phone number, email address, residence, associated IP address, economic situation, education, occupation or language. Another well-known service is *Spokeo*, whose free version is reduced to show full name, gender, age, previous cities and states of residency and relatives. More detailed information about the target requires to pay a premium subscription, which is out of our scope. Similar services would be *Fast People Search*, *Nuwber*, *Cubib* or *Peek You*.

The aforementioned services work correctly for the United States, but if we want to apply OSINT to a target that lives in another country, the use of *Yasni* is more appropriate. However, the results obtained are links related to social networks, addresses and personal contacts, education, and miscellany.

Genealogy services like *Family Search*, *Family Tree Now*, *GENi*, or *True People Search* cover another point of view in searches by providing kinship information. Discovering the family links of our target broadens the amount of information we can unveil, in this case indirectly.

F. LOCATION TECHNIQUE

Researching the locations that our target frequents can give us indications of his/her habits and context. It is also interesting to know the geographic location of a company or the place where an event occurred. In this sense, images, addresses and GPS coordinates are worthwhile data to obtain. TABLE 7 shows some services which are particularly designed to these purposes.

Google Maps, *Wikimapia* or *Bing Maps* are well known sites to find out locations from GPS coordinates. On the other hand, it is also possible to reversely get such information from a location name at *GPS Coordinates*.

TABLE 7. Utility of the OSINT services belonging to the location technique.

Location OSINT service	URL	Main output
Google Maps	`google.com/maps`	Locations from GPS coordinates
Wikimapia	`wikimapia.org`	
Bing Maps	`bing.com/maps`	
GPS Coordinates	`gps-coordinates.net`	GPS coordinates from location
Historic Aerials	`historicaerials.com`	Historic images of the past
Terra Servers	`terraserver.com`	
Land Viewer	`eos.com`	

TABLE 8. Utility of the OSINT services belonging to the IP address technique.

IP address OSINT service	URL	Main output
IP Location	`iplocation.net`	Location, domain and ISP
ViewDNS	`viewdns.info`	Technical network-based information
That's Them	`thatsthem.com/reverse-ip-lookup`	Individual or company information
I Know What You Download	`iknowwhatyoudownload.com`	Torrent files

Note that the images offered by the commented services are continuously updated. However, we could be interested in retrieving old images of past situations. *Historic Aerials*, *Terra Servers* or *Land Viewer* incorporate historic imagery functionalities to precisely discover past and outdated views of locations.

G. IP ADDRESS TECHNIQUE

IP addresses are obtained from cyberattack investigations, email addresses or connections over the Internet. They are also crucial for digital forensic analysis in order to collect as much information as possible from an incident. TABLE 8 summarizes some services which facilitate these tasks.

The service *IP Location* obtains, from a given IP address, high-level aspects such as location (latitude and longitude), country, region, city, domain name or ISP (*Internet Service Provider*). If we are interested in specific facts, the website *ViewDNS* provides more technical information apart from the IP location. In particular, it includes services for displaying registration information about the associated domain name, showing additional domains hosted on the IP address, discovering common ports that may be open and services running on them, or seeing the network path from *ViewDNS* to the target IP address and analyze associated networks, routers, and servers.

Nevertheless, the previous resources provide data that is not sensitive or personal in nature. On the contrary, *That's Them* does offer interesting information about people, home addresses, companies, or emails addresses related with the given IP address.

Another powerful service providing personal information is *I Know What You Download*. This service monitors online torrents and discloses the files associated with any collected IP addresses. The files downloaded by our target could reveal really sensitive information about his behaviour or interests.

H. DOMAIN NAME TECHNIQUE

A typical point of interest in OSINT investigations are web pages. They can reveal interesting information about our target, specially whether we are dealing with a person or a company. It is worth noting that the majority of techniques which are explained for IP addresses are also suitable in this context. In addition to them, we can highlight some other services as presented in TABLE 9.

DNS Trails extracts DNS records, but also identifies the number of additional domains that are related to the encountered results. To this extent, it is a very helpful way to find relationships and connections. *Whoisoly* also shows a cross-reference view from the owner name, address, telephone number or email address.

Another powerful service is *Wayback Machine*, which periodically makes backups of many websites from the whole Internet. This allows an investigator to analyze the evolution and changes of a website, being able to see it for particular screenshots dated in time.

Furthermore, it is possible to visualize domain connections through *Visual Site Mapper* or *Threat Crowd*. Checking DNS and mailservers is also useful by visiting *Whois*, which also offers a ping functionality for checking the connectivity and a traceroute functionality to study the data path to the given domain. There are also services like *Alexa* and *SimilarWeb* which calculate traffic statics and others like *FindSubdomains* which search for subdomains.

VI. OSINT TOOLS

A manual use of some techniques would be enough for basic searches. Unfortunately, using a few services might not be effective for challenging investigations. In this sense, the potential of OSINT lies in using as many services as possible in a concatenated fashion. Following the workflows repeatedly will extend the available information to put all the pieces of the puzzle together. However, it is not practical for

TABLE 9. Utility of the OSINT services belonging to the domain name technique.

Domain name OSINT service	URL	Main output
DNS Trails	`securitytrails.com/dns-trails`	DNS records and related domains
Whoisoly	`whoisology.com`	Personal or company information
Wayback Machine	`web.archive.org/web`	Backups of websites
Visual Site Mapper	`visualsitemapper.com`	Map of subdomains
Threat Crowd	`threatcrowd.org`	
Whois	`who.is`	Registration info and DNS records
Alexa	`alexa.com`	Traffic statics
SimilarWeb	`similarweb.com`	
FindSubdomains	`findsubdomains.com`	Subdomains

TABLE 10. Main features of the selected OSINT tools.

OSINT tool	Input				Output	Extensibility	Interface	Platform	Other feature
	Identity data	Network data	File data	Selectable data source					
FOCA	✗	Domain	File name, Folder	Google, Bing, DuckDuckGo	Identity info, Network info, File info	✗	Stand-alone program	Windows	Server discovery module
Maltego	Personal information, company, community	Domain	File URL	✗	Identity info, Network info, File info	Custom transforms	Stand-alone program	Linux, Windows, MAC	Location, Auto input/ output refeed, Results in oriented graph
Metagoofil	✗	Domain	File type	✗	Network info, File info	✗	Command line	Linux, Windows	Option to narrow results
Recon-NG	Personal information	Domain	✗	Several	Identity info, Network info, File info	✗	Command line	Linux	Location, Modules for discovery and exploitation
Shodan	Country, City, Keyword	Operating system, IP Address, Port, Host name	✗	✗	Network info	✗	Web interface	Online	Location, Webcam captures
Spiderfoot	Email, Real name, Phone Number	Domain, IP Address, Subnet, Host name	✗	Several	Network info	Custom modules	Web interface	Linux, Windows, MAC	Different types of scan, Results in oriented graph
The Harvester	Company	Domain, DNS server	✗	Several	Identity info, Network info	✗	Command line	Linux, Windows, MAC	Results in reports, Option to narrow files and results
IntelTechniques	Personal information, company, community	Domain, IP Address	File name, File type, File URL	Several	Identity info, Network info	✗	Web interface	Online	Location, Public records, OSINT virtual machine

the end user to manually combine several OSINT techniques and their associated services. Such a tedious task would entail lengthy research processes.

For this purpose, researchers and developers have implemented more precise tools for applying OSINT techniques automatically and gathering better quality information from many different sources, implementing several workflows internally and, as a consequence, obtaining further rewarding information and better inferences.

TABLE 10 presents the main features of the most popular and relevant OSINT tools today. We indicate the type of inputs and outputs they allow, the capability of including custom functionalities, the type of user interface, the platform of functioning and other interesting miscellany features.

Nevertheless, there are a lot of OSINT applications in the literature which can be accessed at *OSINT framework*.[4]

A. FOCA

The main contribution of *FOCA*[5] (*Fingerprinting Organizations with Collected Archives*), designed by *ElevenPaths*, is the extraction and analysis of the metadata present in electronic documents. This application can be used for both local files present in our computer and external documents that are downloaded from a specified webpage using three different search engines (*Google*, *Bing*, and *DuckDuckGo*). *FOCA* considers a wide variety of formats such as Microsoft Office, PDF, Open Office, Adobe InDesign, SVG files, etc.

This application extracts the hidden information of the files and processes them to show the user relevant aspects. Some of the details that are discovered with this procedure are the name of computers related to the documents, the location where the documents were created, operating systems used,

[4] osintframework.com

[5] https://www.elevenpaths.com/es/labstools/foca-2

real names and email addresses of related users, data about the servers, date of creation of the documents, range of IP addresses of internal networks, etc. As a result, a network map can be drawn based on the extracted metadata to recognise the target.

FOCA additionally includes a server discovery module to complement the metadata analysis of documents. Some techniques used in this tool are: (i) *Web Search* for searching hosts and domain names through URLs associated to the given domain; (ii) *DNS Search* for discovering new hosts and domain names through the NS, MX and SPF servers; (iii) *IP Resolution* for obtaining the IP addresses of encountered hosts through the DNS; (iv) *PTR Scanning* for finding more servers in a discovered network segment; (v) *Bing IP* for extracting new domain names associated to encountered IP addresses.

This tool is usually used in the security sector as it allows pentesting a company. In fact, it is able to output very good results because companies do not usually clean metadata from files that are uploaded to the network.

B. MALTEGO

Maltego[6] is a well-known application that automatically finds public information about a certain target within different sources (DNS records, Whois records, search engines, social networks, various online APIs, files metadata, etc). The relationships between the found items of interest are represented in the form of a directed graph for its analysis. This tool defines four main concepts:

- **Entity**: is a node of the graph representing the discovered piece of information. Some default entities are real name, email address, username, social network profile, company, organization, website, document, affiliation, domain, DNS name, IP address, and so on. Furthermore, we could also define custom entities for our specific investigation.
- **Transform**: is a piece of code which is applied to an entity to discover a new linked entity. For example, the transform "*To IP Address*" which resolves a DNS name to an IP address, could be applied to a domain name entity "`um.es`" to create a new IP address entity "`155.54.212.103`". Recursively, we would continue applying more transforms, propagating the process of search. Apart from default transforms, it is also possible to implement and include custom ones for more specific purposes.
- **Machine**: is a set of transforms that are defined together to be executed in order to automate and concatenate long processes of search.
- **Hub Item**: is a group of transforms and entity types used to allow users of the community to reuse them. By default, Maltego implements the hub item called "Paterva CTAS" which contains the entities, transforms and machines maintained by official developers. In addition, it is possible to create and install third party hub items.

C. METAGOOFIL

Metagoofil[7] works similarly to FOCA. It is a gathering tool which downloads public files found in a target domain or URL and extracts their metadata to output knowledge. It generates a useful report for pentesters with usernames, real names, software versions, and servers or machine names. It can also find further documents that could contain resources names.

Although it is a command line functionality, some interesting options in favor of OSINT investigations are permitted. Apart from specifying the target domain or the local folder to analyze, *Metagoofil* allows filtering filetypes (pdf, doc, xls, ppt, odp, ods, docx, xlsx, pptx), narrowing down the results to search and the number of documents to download, determining the working directory where downloaded files are saved, or selecting the file to write the output.

D. RECON-NG

Recon-NG[8] is a web recognition framework similar to Metasploit.[9] It presents a command line interface that allows one to select a module to use, which is essentially an OSINT resource. Then, we set some parameters if necessary and launch the process. The results of the searches are continuously saved in a workspace which in turn feeds next rounds of the process.

This tool includes several independent modules that implement different functionalities. For example, the modules *Bing Domain Web* and *Google Site Web* search in *Bing* and *Google* search engines respectively for hosts connected to the domains of the workspace; *PGP Search* scans the stored domains to find email addresses associated with public PGP keys; *Full Contact* gathers users and corresponding social networks profiles in its database considering stored contacts; or *Profiler* searches for additional online services that possess accounts with the same user names as those in the workspace.

Recon-NG is continuously agglutinating in a local database all the obtained information. In this way, the user directs the research by selecting the indicated module and the tool automates the generation of knowledge from there. The system scales remarkably for complex investigations.

E. SHODAN

Shodan[10] is a search engine that provides public information of Internet-connected nodes, including IoT devices. This includes servers, routers, online storage devices, surveillance cameras, webcams or VoIP systems, amongst others. The recollection of data is made through protocols like HTTP or

[6] https://www.paterva.com/web7/buy/maltego-clients.php

[7] https://github.com/laramies/metagoofil

[8] https://bitbucket.org/LaNMaSteR53/recon-ng/wiki/browse

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

[10] https://www.shodan.io

SSH, allowing the user to search by IP address, organization, country name or city.

This tool is mainly used for network security (to find devices exposed to the outside or detecting vulnerabilities of publicly available services), internet of things (to monitor the growing usage of smart devices and their location in the world geography), and tracking ransomware (to measure the infection provoked by this type of attack). It allows downloading the results in JSON, CSV or XML formats, as well as generating user-friendly reports.

In addition to the mentioned functionality, there are two premium services, namely: *Shodan Maps* (`maps.shodan.io`), permitting investigations based on locations, and *Shodan Images* (`images.shodan.io`) displaying collected images from public devices.

F. SPIDERFOOT

Spiderfoot[11] is another reconnaissance tool that automatically goes through lots of public data sources to compile information. Our input could be an IP address, subnet, domain name, e-mail address, host name, real name or phone number. The results are represented in a graph of nodes with all the entities and relationships found.

Depending on the type of input introduced, this tool autonomously selects the modules (equivalent to Maltego transforms) to activate for a more effective reconnaissance. Moreover, it also considers the level of search selected by the user. *Spiderfoot* offers four types of scans: (i) *Passive* collects as much information as possible without touching the target site, avoiding being unveiled by the target; (ii) *Investigate* conducts a basic scan in order to find out target's maliciousness; (iii) *Footprint* identifies the network topology of the target and gathers information from the web and search engines, sufficient for standard investigations; and (iv) *All*, which is advisable for detailed investigations, despite taking a long time to complete, as it consults absolutely all possible resources related to the target.

This tool could be used to launch penetration tests to reveal data leaks and vulnerabilities, red team challenges, or to support threat intelligence. In addition, it is worth noting that it is possible to program custom *Spiderfoot* modules.

G. THE HARVESTER

The Harvester[12] allows the collection of public information related to a domain or company name through search engines. In particular, it is capable of listing emails and host names of the company, as well as subdomains, IP addresses and URLs related to the domain. It also permits user-friendly HTML or XML representations of the results. This resource is used in the early stages of a penetration test.

This tool is managed from the console and implements two options when scanning our target website. On the one hand, *The Harvester* represents the original script which actually provides the list of related email addresses, whereas, on the other hand, *EmailHarvester* improves the procedure by digging deeper for better results.

H. INTELTECHNIQUES

IntelTechniques[13] is a tool, created by Michael Bazzel, which offers hundreds of online search utilities grouped by technique.

When using it, the investigator selects the services to be used and this tool automatically creates the associated query links. Afterwards, the user can enter them in the browser to launch the queries. However, the visualization and collection of the information is still manual.

In spite of the fact that it does not implement an automatic integration of services, we have considered *InterTechniques* as a OSINT tool that facilitates the launch of searches to a wide range of services from a centralized platform.

Unfortunately, this tool ceased to be free and blocked its open access as of July 2019 due to constant cyberattacks.

I. OSINT TOOLS COMPARISON

Depending on the user needs (see TABLE 10), some tools will be more suitable than others for a given task.

Thus, if we intend to extract **hidden information from files**, *FOCA* and *Metagoofil* are specific tools designed for this purpose. In particular, the first product seems to be more complete, mature and powerful than the second one. *FOCA* presents additional functionalities, apart from the metadata analysis of files, to complement the hidden information. As a result, it is able to infer more knowledge about the target.

Yet, if we are looking for **network information**, *Shodan*, *Spiderfoot* and *The Harvester* are recommended options for this certain task. On the one hand, we would suggest *Spiderfoot* to analyze the topology of the target and retrieve internal (but public) information about the target organization. On the other hand, we would complete the results with *Shodan* to include specific information about IoT devices, surveillance cameras, webcams, VoIP systems, or smart services in general.

Last but not least, if the aim of the search is to gather **as much information as possible** for a given input, the resources *Recon-NG* and *Maltego* are the more complete ones and will return diverse data and relationships. The first one contains lots of modules and interacts with a local database that scales during the investigation, being an ideal framework to carry out pentestings, phishing and social engineering attacks prevention, or even the profiling of a person. On the contrary, if we want to avoid the command line and opt for a more user-friendly interface, *Maltego* is a good alternative for OSINT activities. It implements automated inference processes with transforms that raise the scope of the original search. Moreover, it is extensible with custom discovery procedures.

[11] https://www.spiderfoot.net

[12] https://github.com/laramies/theharvester

[13] https://inteltechniques.com

Despite the fact that the above described comparison has been made according to the desired output, in practice the user will be restricted by the available input and the data type accepted by the chosen OSINT tools. Finally, note that these tools are complementary and mutually non-exclusive, meaning that a deep and thorough OSINT investigation could profit from several of them at the same time. Although some of them may produce similar results for a given search, there can always be details found by a particular tool that are not obtained by others.

VII. INTEGRATION OF OSINT IN CYBERATTACK INVESTIGATIONS

The implementation of mechanisms for detection of and response to cyberincidents is an obligation today. Companies and organizations, which are increasingly exposed on the Internet, invest in cybersecurity to protect their assets against criminals. Therefore, it is remarkably important to manage threats and incidents against information systems effectively.

Cyberdefence is not only the deployment of technical solutions such as firewalls, IDSs (*Intrusion Detection Systems*), IPSs (*Intrusion Prevention Systems*), SIEMs (*Security Information and Event Management*) or anti-viruses to avoid known threats, but also the implantation of cyberintelligence to extract and analyze traces, patterns and conclusions from the incidents. In fact, the continuous cycle of extracting and sharing evidences, relationships, and consequences of incidents is known as threat intelligence [65]. It complements the traditional defence mechanisms with up-to-date information and highly improves the protection of the infrastructures, the management of the hazards and the effectiveness of the responses [41].

Moreover, the information that is typically used for forensics and investigations is merely technical. However, the traces left by a cyberattack contain valuable information that should not only be contrasted with repositories of incidents [66], but also with social networks, forums, media, technical and governmental documents and other digital public sources. These open sources contribute with semantic information in the analysis, which result to be interesting for computing and reasoning more complex and far-reaching inferences. Note that cyberattackers use the Internet for their illegal actions (hacking, phishing, denial of service attacks, botnets, identity theft, intrusions, etc.), but also for personal reasons. In this sense, OSINT can be used to connect all those points.

Several works applying OSINT to cybersecurity focus on proposing defensive improvements when facing threats. On the contrary, very seldom they seek the identification of cyberattackers. OSINT is a source of knowledge that could support the investigation of a cyberattack by going from the smallest details of the malicious action to the root of the problem. This last challenge is not new, since it is traditionally known as the attribution problem [67]. Concretely, OSINT would allow us to understand the motivation of the cyberattack, to guess the procedure and to ultimately profile the perpetrator.

The suggested application of OSINT is illustrated in FIGURE 3. Note that several methodologies and models have been proposed to define the detection maturity of an organization, which is crucial to extract evidences from a suffered cyberattack. Nonetheless, there is a lack of standards to represent taxonomies and ontologies in this field [68], thus we propose a modified version of Ryan Stillions' DML model [69] to exemplify this section. However, another cyberthreat detection scheme could be used to show the application of OSINT in a similar way.

The DML model represents in a hierarchical way different levels of abstraction in the detection of cyberattacks. A company that does not invest in cybersecurity will only be able to reach the lowest steps in the stack. On the contrary, an organization technically skilled in cyberdefence may interpret more complex facts, that is, to ascend to levels with more abstraction.

While the lower levels can be easily covered, the challenge lies in reaching the higher layers. To this end, we suggest applying OSINT as a source of intelligence that feeds on the most basic evidence to arrive at more robust facts:

1) Firstly, we assume that it is possible to cover levels DML-1 and DML-2. The first one, *Atomic indicators of compromise (IOC)*, is composed by details as simple as a string in a modified file, the value of a memory cell or a byte transmitted through the network, which have very low value on their own, but together form the next level. The *Host and Network Artifacts* layer is built upon the indicators observed during or after the cyberattack such as IP addresses, domain names, logs, transactions, hash values, or file manipulation details.
   As this type of data resides in the affected information systems, in our framework it is considered as an input for the collection of associated information in open sources (see SECTION V for more details about OSINT collection). Therefore, the extraction of these traces is the starting point of an OSINT process.
2) Next we have from level DML-3 to level DML-6. The third level *Tools* consists in detecting the transfer, presence and functionality of the tools used by the attacker. The following level *Procedures* is covered if one is able to enumerate the steps performed during the incident. The fifth level *Techniques* extracts how the attacker has specifically performed the various phases of the attack. And the last level here, *Tactics*, is a more abstract concept that takes into account the levels discussed above and derives knowledge by analyzing a set of activities in time and context.
   In this case, the information reveals details about the execution of the cyberattack. Such data highly enriches the analysis phase of the OSINT cycle. The patterns derived from this data, as well as the correlation with other cases already stored, allow us to have a more intelligent and comprehensive analysis. In fact, these

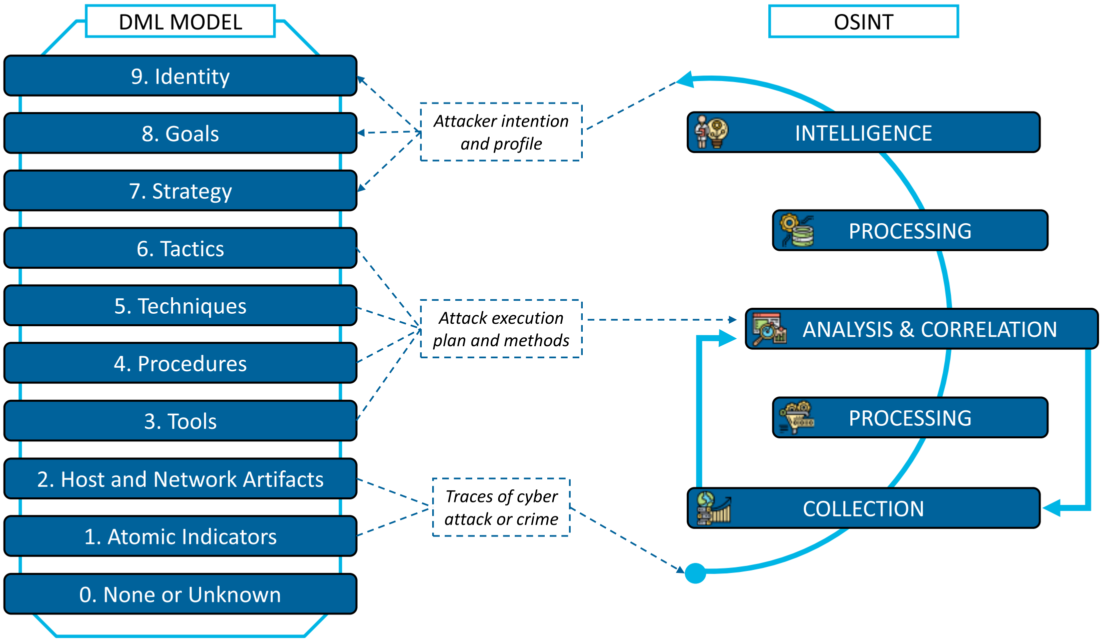

FIGURE 3. OSINT integration with DML model to address the attribution problem.

conclusions should be integrated in conjunction with the results obtained in the collection phase. In this way the exploration through the network is refined, narrowing the investigation towards the final objective.

3) Finally, the continuous gathering and analysis process of OSINT generates valuable information to which knowledge-extraction techniques are applied. The knowledge extracted with OSINT from level DML-1 to DML-6 would allow us to reach the highest levels, that is, from DML-7 to DML-9. The seventh level, *Strategy*, refers to a high-level description of the planned attack of the cybercriminal to complete his/her purposes. The eighth level, *Goals*, are the specific objectives of the attacker and express the real motivation of the action. At the top we find the *Identity* level, which is essentially the name of a person, an organisation or even a country which is responsible of the malicious actions. As it is extremely difficult to find that detailed information, the connection with other cyberattacks and the similarity with other events can support the relative attribution [67]. That is, completing the investigation of the current case with additional information about other incidents apparently caused by the same actor brings us closer to the absolute identification of the cyberattacker.

This application of OSINT represents an innovative line of action to fight against cyberthreats. The challenge resides in implementing effective mechanisms of collection and intelligent analysis procedures to extract those high-level details that can not be directly extracted from malicious actions. Such details are the most complicated pieces of information to achieve, as they have a very high degree of abstraction that are long away from the technical details. That is why it is smart to look to open sources for any relationship or pattern that leads us to discover more about the context and originators of an incident. OSINT is the key piece that was missing in the gear to profile cyberattackers and to improve the detection of sophisticated attacks [70] thanks to the consideration of high-level behaviour aspects from DML-3 to DML-9.

VIII. OSINT IN COUNTRIES AND STATES

OSINT is not only beneficial in the private sector, but also represents a resource of public interest in governments. In this regard, in SUBSECTION VIII-A we discuss that OSINT is not a paradigm designed for paranoid analysts or computer geeks, but indeed has an enormous benefit in the cyberdefence national system [71]. Likewise, in SUBSECTION VIII-B we observe that official authorities do not only get profit from OSINT results for internal tasks, but indirectly make the application of OSINT easier for third parties. In fact, they become an agent that generates large amounts of data accessible to everyone. In this sense, governments are a double-edged sword which benefit from OSINT but at the same time they contribute to feed the Internet with really valuable, and sometimes even sensitive, information.

A. INTERNAL STATE AFFAIRS OPERATIONS

Intelligence Agencies have been traditionally associated with the labour of Law Enforcement Agencies (LEAs) and Military Bodies. In the same way, OSINT is considered nowadays as an important key of classified investigations and secret operations in state affairs [5]. To some extent, one could safely argue that the exploitation of OSINT can provide critical capabilities for LEAs to complement and enhance their counterintelligence departments in the investigation and strategical planning to fight against crime [72].

As far as we were able to explore in the official websites, reports and documentation, government organizations seem to implement internal mechanisms which basically consist in gathering raw information and transforming it into useful knowledge, leveraging OSINT mechanisms [73]. In a representative way, we could mention the *U.S. Federal Bureau of Investigation (FBI,* `fbi.gov`*)*, *U.S. Central Intelligence Agency (CIA,* `cia.gov`*)*, *Canadian Security Intelligence Service (CSIS,* `canada.ca/en/security-intelligence-service`*)*, *European Union Agency for Law Enforcement Cooperation (EUROPOL,* `europol.europa.eu`*)*, *North Atlantic Treaty Organization (NATO,* `nato.int`*)*, *United States Department of Army (DA,* `army.mil`*)*, *U.S. Department of Defense (DoD,* `defense.gov`*)*, *U.S. National Security Agency (NSA,* `nsa.gov`*)* or *European Defence Agency (EDA,* `eda.europa.eu`*)*, amongst others.

In this scenario of uncertainty, we have decided to particularly investigate the case of Spanish LEAs, for affinity, to demonstrate that official organisms internally indeed apply OSINT. As a result of this thorough inspection, we can emphatically confirm that it is not easy to find clear evidences of the application of OSINT by the state forces. The confidentiality of this type of agencies makes it difficult to discover their internal operating mode and the impact of OSINT in their current investigations. Nevertheless, as a consequence of the deep search, we have some subtle findings that confirm that OSINT is currently used by Spanish LEAs:

- Back in 2007, the director of the CNI (i.e., Spanish National Intelligence Agency) said[14] that open sources were "*fundamental to the elaboration and work of Intelligence*"
- CIFAS (i.e., Spanish Military Intelligence Agency) also seems to use OSINT as a way of obtaining information. We have found some slides that confirm this, dated as early as in 2008, which are uploaded in the Spanish Defense Staff website.[15]
- In 2010, when the director of the CNI announced[16] the creation of an ethical code for special agents, he also insisted on the fact that modern intelligence was not just based on physical presence, as today "*you might get more information sitting on a computer, exploring messages from the bad guys*".
- More recently, in 2017, the Spanish Ministry of Defense opened a public call[17] for the contract called "*Development of OSINT tool based on IDOL HAVEN platform*".
- In the present, the Spanish Army is designing a new model called *Brigade 2035* which incorporates innovative technological advances for enhancing operations. In this project,[18] one of the defined combat functions is *Intelligence*, which clearly states OSINT as a key responsibility: "*Other facilities of growing importance will be open source obtainment (including social networking)*".
- The Spanish Ministry of the Interior has published in the Annual Recruitment Plan for 2019[19] some investments in "*systems for obtaining OSINT in the cyberspace*".

Bearing in mind all these facts, it seems that currently OSINT is indeed relevant in the internal affairs of Spain. Analogously, we could also highlight that European Union state members are also highly developed in OSINT [74].

B. OPEN DATA POLICIES AND TRANSPARENCY

OSINT depends on the public data available on the Internet, among other sources, to be effective. In this regard, apart from social networks and other open data sources, there are also authoritative and official sites maintained by state institutions around the world where public information is published and, therefore, openly available.

The Open Data Barometer (ODB)[20] is a global ranking system designed by the World Wide Web Foundation that measures the readiness, implementation and impact of countries' open data policies. In Figure 4 is shown the scores of latest full edition.[21]

As we have already done in the previous subsection, we study the specific case of Spain for affinity. In fact, regarding the aforementioned ODB report, Spain is ranked in the 11th position. Besides, according to the European Data Portal and its official reports[22] about Open Data maturity across Europe, Spain is one of the most advanced countries in transparency and open data. It has been in first or second position in the ranking of Open Data Maturity in the last four years. As it is stated, the Spanish Government has promoted more than 160 open data initiatives and has over 23,800 public information catalogues. For example, the Open Data Initiative of the Government of Spain[23] is a clear proof of how Spain encourages transparency. OSINT could benefit from that, but it should deal with aggregated and statistical information by linking it and inferring new knowledge.

There are also anonymized databases that, a priori, would not be useful for OSINT because they lack the value to produce intelligence. These so-called anonymous datasets do not break the link between the data and its owner, apparently. Recently, an algorithm [75] has been published allowing 99.98% of Americans to be unequivocally identified from public data. In particular, it is enough to have 15 parameters related to medical, behavioral and socio-demographic

[14] https://www.elconfidencialdigital.com/articulo/vivir/CNI-califica-fundamental-abiertas-contradice/20071023000000049386.html

[15] http://www.emad.mde.es/Galerias/EMAD/novemad/fichero/EMD-CIFAS-esp.pdf

[16] https://www.lavanguardia.com/politica/20100624/53951898847/el-director-del-cni-anuncia-un-codigo-etico-para-los-agentes-secretos.html

[17] https://contrataciondelestado.es/wps/wcm/connect/ff96fa82-7fd6-40bd-be5b-36ef3fd4e65b/DOC_CN2017-498874.pdf?MOD=AJPERES

[18] www.ejercito.mde.es/en/estructura/briex_2035/principal.html

[19] http://www.defensa.gob.es/Galerias/gabinete/ficheros_docs/2019/PACDEF_2019_Documento_Pxblico.pdf

[20] https://opendatabarometer.org

[21] https://opendatabarometer.org/4thedition

[22] https://www.europeandataportal.eu/en/dashboard#2018

[23] https://datos.gob.es/es

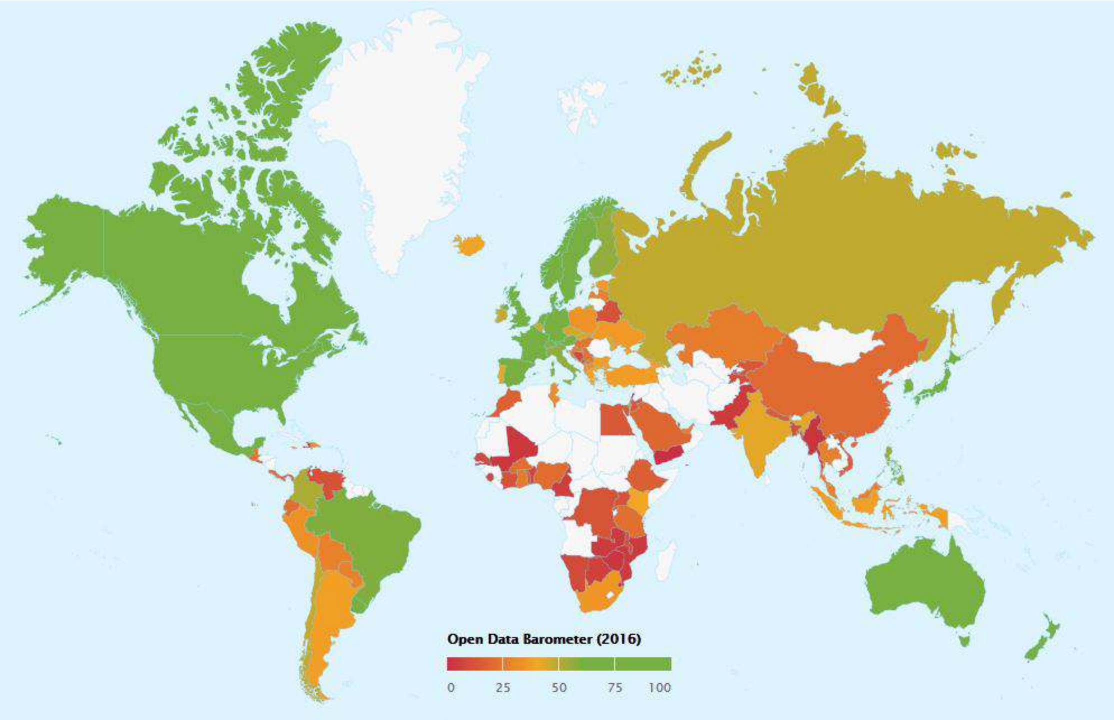

FIGURE 4. Transparency scores by the 4th edition of Open Data Barometer.

information such as marital status, sex or the zip code of their home. Therefore, OSINT could again be used to re-identify people collected in anonymized databases.

On the contrary, there are also governmental platforms which are actually not anonymized. For instance, the Spanish Ministry of the Treasury, the Spanish Ministry of the Interior or the Spanish Ministry of Defense usually publish documents with personal information (“`site:hacienda.gob.es filetype:pdf intext:dni`”, for example). In the same way, this could be also applied to Spanish Autonomous Communities websites. Moreover, Europe has a public data platform[24] too, where we could find a lot of public information. For instance, in the context of foreign policy and security, an updated list of financial sanctions is presented in the “*European Union Consolidated Financial Sanctions List*” document. In particular, it reveals personal information about individuals, groups and entities.

All the aforementioned facts demonstrate that governments worldwide are adopting strong Open Data policies. As a direct consequence, the amount of objective data available on the Internet is rapidly increasing. OSINT should, in addition to other open sources of information, take advantage of this powerful opportunity to collect, analyze, link and infer knowledge from reliable and official sources. In this scenario, and according to the ODB, countries such as United Kingdom, Canada, France, United States, Korea, Australia, New Zealand, Japan, Netherlands, Norway, or Brazil are real OSINT goldmines with very similar characteristics to those commented for Spain.

IX. OPEN CHALLENGES AND FUTURE TRENDS

The review carried out on OSINT shows that there is already a substantial amount of work in the topic. Numerous techniques and tools have been developed up to now. However, there are some gaps and limitations in this field to continue exploiting the offered opportunities. It is necessary to make more sophisticated solutions applicable to uncontrolled scenarios of the real world. We have spotted some challenges that, as far as we know, are open nowadays and should be faced by the research community in the next future.

A. AUTOMATION OF THE GATHERING PROCESS

The greater the amount of information collected, the more likely it is to create inferences and relationships. However, the quantity of public data available today is enormous and can not be collected in a manual way [76]. Although OSINT techniques (Section V) and tools (Section VI) are already a big step forward in this direction, most of them are still largely dependent on the end user. In this sense, it would be appealing to incorporate more sophisticated techniques. We highlight current big data techniques such as Web crawling or Web scraping [77] as potential paradigms to automate and improve the OSINT exploration of high volumes of open data.

An important aspect of the recollecting process is the propagation of the search. The results obtained with searches should refeed the following rounds of gathering. In OSINT it is really powerful to extract pivots permitting

[24] http://data.europa.eu/euodp/en/data

the concatenation of outputs as new inputs for propagation. This recursive method increases the scope of research and is closely related to the analysis process that we will discuss next.

B. ENHANCEMENT OF THE ANALYSIS AND KNOWLEDGE EXTRACTION PROCESSES

The interpretation of the recollected open data is a key point in the OSINT procedure. Extracting the essence of the scraping results, making relationships between separated pieces of information, or inferring conclusions that are not explicitly exposed increases the quality of the results. Indeed, the recursive integration with the propagation of further rounds of investigation is enhanced by means of better inputs.

However, as far as we know, OSINT analysis is not implementing intelligent mechanisms today. The existing tools are limited to throwing all the information found and its explicit relationships. On the contrary, the analysis process should incorporate semantic analysis, study of patterns, correlation with other events, occurrences or datasets.

Fortunately, modern data mining techniques [78] such as Natural Language Processing, Social Network Analysis, Machine Learning or Deep Learning are actually designed to solve this type of challenges. A proper selection of algorithms in this field of knowledge will make the difference between the current static analysis and the future reasoned processing [79].

Ideally, the OSINT of the future should be able to provide the end user with the specific piece of information he/she is searching, as well as to return convincing answers in investigations. The original search would also have, not only direct inferences, but also indirect and not explicit relationships.

This challenge builds the path between the Second Generation and the Third Generation of OSINT. As it is presented in [1], the Second Generation started with the rise of Internet and Social Media, and the challenges were "*technical expertise, virtual accessibility and constant acquisition*". In contrast, the evolution to the Third Generation is supposed to appear nowadays and will have to include "*direct and indirect machine processing of data, machine learning, and automated reasoning*".

C. INTEGRATION OF SEVERAL OPEN DATA SOURCES

OSINT activities should consult as many sources as possible in order to cover the widest possible spectrum. It is not a good idea to focus our research on a single social network or a specific forum. In this sense, success lies in combining data sources to obtain the best possible results. This means that the system has to normalize the available information, which is typically unstructured, in order to perform an effective analysis and correlation. As a result, it is important to discard repeated items. In fact, the different OSINT techniques and tools explained in this paper are actually applying such sitting to gather the knowledge related to the target.

On the other hand, the real challenge is to incorporate, not only several data sources, but different types of data sources [80]. Apart from data extracted from the Internet, Dark Web and Deep Web, the OSINT workflow should also consider information collected face to face, with social engineering, or with citizens collaboration. Any piece of information which is interesting to our investigation has to be used in order to achieve the next milestone of the search. Additionally, it is a must the implementation of truth discovery processes for those cases when information from different data sources is contradictory [81].

D. FILTERING OUT IRRELEVANT DATA AND MISINFORMATION

Due to the huge amount of data publicly available, an OSINT process needs to be capable of distinguishing the relevance of each piece of information, discarding data which do not add quality to the investigation [82]. A researcher cannot focus on exploring the details of an entire website, reading a multi-page news item or analyzing a complex government document. On the contrary, OSINT research needs to extract keywords which actually provide value and reveal knowledge about our target. The piece of information we are interested in may not be explicitly posted, and the challenge would be to extract the essence of the data source we are scrutinizing. At the same time, the precise terms extracted serve as pivots to create new paths of exploration.

Furthermore, it is crucial to detect misinformation that would corrupt the results [83]. By nature, the Internet is subjective and the majority of the content has no guarantee of being reliable and official. The OSINT community has to determine whether the increasing reliance on open source data is still combined with the sources validation, which represents a primary requirement and priority [84]. That untrue information can divert our search, leading to erroneous results or far from our real objective. For that reason, it would be interesting to analyze not only the objective information, but also the false information with the aim of extracting intelligence.

This problem will be present in real-life research. The data sources where we will find more valuable information about suspects will be in forums and social networks. In these sites, the investigator has to deal with opinions, subjective publications, and personal preferences whose veracity is questionable [85]. Profiling of persons who in reality do not represent a threat (false positives) could provoke discriminatory and unfair attitudes that could affect the victims.

E. EXTENSION ACROSS THE WHOLE WORLD

One of the main drawbacks of many of the existing OSINT resources is that they only function for specific countries, reducing their profiling capability to a constrained group of people belonging to a few nationalities. However, OSINT should be a universal technique to tour all the corners of the Earth instantly without discriminating zones of the cyberspace. Thus, interoperability is a desirable property to be considered in OSINT design as it would increase, not only the scope of the searches, but also its usage by end users.

Ideally, a good OSINT service or tool should not distinguish between countries and take each research as a global task, without borders. The OSINT workflow should combine points of information across the world and correlate those distributed data sources. In fact, although the relationships between search zones could be done by hand, the real challenge lies in OSINT applications implementing these *jumps*.

In addition, the globalization of the process would not leave aside appealing open data sources from different territories which actually could fill the gaps we need to address in our investigation. In Spain, for instance, we use tools that are designed in (and for) foreign countries. However, there are not OSINT solutions which include Spanish public repositories in the gathering phase (as government open data platforms could be). In this sense, we are not fully benefiting yet from the goldmine that supposes being one of the most transparent countries in Europe.

A generic and flexible implementation is specially useful for *nomad targets* in whom mobility is part of their daily lives. Say that the investigated target is a person who has lived stages of his life in several countries, or companies which have headquarters on several continents, or even criminals who change their location to make it more difficult to pursue them. In these cases, a static search in a particular country would leave a lot of information uncollected and a lot of clues unanalyzed.

F. AWARENESS OF PRIVACY, ETHICAL AND LEGAL CONSIDERATIONS

From an ethical point of view, OSINT must respect the user's privacy so as not to harm his private life, as well as the privacy of his family, friends and co-workers. The fact that the information is publicly accessible does not mean that it is not sensitive. Knowing the personal preferences and tastes of the target can perpetrate in his privacy. Revealing political thoughts can have fatal consequences in certain places. Communicating a sexual orientation can be potentially life threatening in certain countries. Knowing religious beliefs can lead to criminal convictions in specific territories. Thus, the open source information has to be handled carefully, for legitimate purposes, in the interests of society.

From the legal point of view, OSINT should be used on the basis of a law and respecting data protection policies. With the advent of the EU GDPR, the regulation concerning the personal data has changed [86]. In this sense, personal data comprise any information which can relate to any citizen. Moreover, different pieces of information, which collected together can lead to the identification of an individual, also constitute personal data, even if the information is encrypted or anonymized [14]. A possible solution to address such challenge is to adapt the design of OSINT tools to embed normative constraints, specially GPDR legal requirements [87]. By definition, OSINT is completely legal due to the public nature of the data sources it uses. Nevertheless, investigators must not publish the gathered personal information, even if it is posted on the web. In addition, the user who applies OSINT cannot fall into the error of trying to impersonate the target in order to find more information. It should also be noted that authentication barriers cannot be broken in order to access the information we are looking for.

In short, the use of OSINT should be restricted to legal activities and non-malicious purposes. In principle, OSINT does not (and should not) violate human freedom and rights, therefore its previously-mentioned techniques and services are legal to this extent [88]. It is a really powerful methodology, but it is also dangerous if misused. Thanks to OSINT, journalists can provide up-to-date, objective and quality news. Human resources managers can get to know the applicants in their job better. Countries' authorities can investigate criminal and terrorist groups. A company can audit its exposure abroad to cyberthreats. However, such openness to the utilization of OSINT techniques to specific categories should be always correctly justified [89].

On the downside, the OSINT end-user could be a delinquent trying to commit a crime. A cracker could profile the target to increase the likelihood of success. A thief could analyze family members to steal from home at the best time. An extortionist could publish the private and personal information of the victim if a ransom is not paid.

Developers have to consider the aforementioned aspects when implementing OSINT tools. In any case, for our sake, the most powerful tools should be only available to LEAs and Intelligence Agencies.

G. BATTLE AGAINST OSINT MISUSE

As already mentioned throughout the previous Sections, the potentialities of the OSINT paradigm are quite broad. In fact, it is indeed possible to take advantage of the open data for cybersecurity and cyberdefence purposes, thus investigating the attackers and/or terrorist groups [90]. Nevertheless, the exploitation of the publicly-available data is prone to abuse. That is, ill-motivated actors may leverage the huge amount of information in order to commit cyber-aggressions, such as cyberbullying, cybergossip and cyber-victimization [91]. Unfortunately, those phenomena are increasingly and alarmingly more frequent on the Web, leading the victims to distress, loneliness, depression, and even to commit suicide in the worst case [16]. In particular, cybergossip is performed by group of people making evaluative comments via digital devices about somebody who is not present. This cyberbehavior affects the social group in which it occurs and can hinder peer relationships, damaging the victim of such process [92].

To this extent, it is important to control that the OSINT techniques and services are used in the correct manner, without harming others' rights and freedom [93]. More specifically, one could think to give different privileges based on end-user category, thus avoiding to grant full-access to the entire spectrum of information. For example, employees may have access to basic information in order to enhance their tasks (e.g., for HR recruitment duties), while government and police forces may explore and investigate more open data (e.g., to hunt a cyber criminal).

Finally, it is important to note that OSINT is enabling new proposals to combat this scourge of cyber-aggressions [94]. In this sense, OSINT misuse is likely to be properly detected actually with OSINT-based tools.

X. CONCLUSION AND FUTURE WORK

The widespread use of forums, social networks, or the media, as well as the large amount of existing data, turn Open Source Intelligence (OSINT) into the next Internet goldmine. The extraction of knowledge from public sources represents a way of resolving existing problems from a different and innovative perspective. Specifically, cybersecurity and cyberdefense can be greatly benefited by the results that this type of intelligence can offer. Therefore, automated OSINT processes should be implemented, capable of taking investigations to all parts of the Internet and extending our mind through the web.

This paper described the status of OSINT today. It revealed that the effectiveness of current works is questionable due mainly to their poor application in real scenarios. In fact, there is a lack of serious approaches for transforming OSINT into a robust and self-managed solution. Nevertheless, we suggest the integration of OSINT into existing cyberdefence mechanisms to move from the atomic technical trails of a cyber incident to the profile of the culprit or the identity of the suspect. The article also presented some OSINT techniques for basic searches and described the most sophisticated OSINT tools nowadays for advanced investigations. Depending on the data available and on the ultimate goal, a proper selection of the most appropriate tool would mark the difference. However, a varied combination of them is actually the key to achieve plausible results.

In the context of Spain, we pointed out some indications which might confirm that Spanish Law Enforcement Agencies and Intelligence Services employ OSINT in their internal procedures. Despite being a confidential aspect of their functioning, OSINT is a crucial element in the context of their investigations. It is worth pointing out that Spain would be a large territory where to research, develop and apply this methodology due to its Open Data maturity. Actually, it is one of the most transparent countries of Europe, according to the European Data Portal.

As future research directions, the article outlined some open challenges related to gathering, analyzing and extracting real knowledge from the immersion of the Internet. Aspects such as misinformation, privacy, and legality will be prominent in the future of OSINT. There is still a long way to go in this area, and to that end the community should address the discussed challenges by including advanced techniques and improving the current performance. The OSINT ultimate goal is to be able to ensure the desired finding for a certain purpose, in an automated and a self-driven way.

REFERENCES

[1] H. J. Williams and I. Blum, "Defining second generation open source intelligence (OSINT) for the defense enterprise," RAND Corp., Santa Monica, CA, USA, Tech. Rep. RR-1964-OSD, 2018, doi: 10.7249/RR1964.

[2] M. Nouh, J. R. Nurse, H. Webb, and M. Goldsmith, "Cybercrime investigators are users too! Understanding the socio-technical challenges faced by law enforcement," in *Proc. 2019 Workshop Usable Security*, Feb. 2019.

[3] A. Powell and C. Haynes, "Social media data in digital forensics investigations," in *Digital Forensic Education: An Experiential Learning Approach*, X. Zhang and K.-K. R. Choo, Eds. Cham, Switzerland: Springer, 2020, pp. 281–303.

[4] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Inf. Fusion*, vol. 28, pp. 45–59, Mar. 2016.

[5] H. L. Larsen, J. M. Blanco, R. P. Pastor, and R. R. Yager, Eds., *Using Open Data to Detect Organized Crime Threats: Factors Driving Future Crime*. Cham, Switzerland: Springer, 2017.

[6] M. Dawson, M. Lieble, and A. Adeboje, "Open source intelligence: Performing data mining and link analysis to track terrorist activities," in *Information Technology—New Generations*, vol. 558. Cham, Switzerland: Springer, Jul. 2018, pp. 1–11.

[7] F. Ali, F. H. Khan, S. Bashir, and U. Ahmad, "Counter terrorism on online social networks using Web mining techniques," in *Intelligent Technologies and Applications*, I. S. Bajwa, F. Kamareddine, and A. Costa, Eds. Singapore: Springer, 2019, pp. 240–250.

[8] J. Jang-Jaccard and S. Nepal, "A survey of emerging threats in cybersecurity," *J. Comput. Syst. Sci.*, vol. 80, no. 5, pp. 973–993, Aug. 2014.

[9] F. Gómez Mármol, M. Gil Pérez, and G. Martínez Pérez, "I don't trust ICT: Research challenges in cyber security," in *Trust Management X*, S. M. Habib, J. Vassileva, S. Mauw, and M. Mühlhäuser, Eds. Cham, Switzerland: Springer, 2016, pp. 129–136.

[10] P. Nespoli, D. Papamartzivanos, F. Gomez Marmol, and G. Kambourakis, "Optimal countermeasures selection against cyber attacks: A comprehensive survey on reaction frameworks," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 1361–1396, 2nd Quart., 2018.

[11] D. Quick and K.-K.-R. Choo, "Digital forensic intelligence: Data subsets and open source intelligence (DFINT+OSINT): A timely and cohesive mix," *Future Gener. Comput. Syst.*, vol. 78, pp. 558–567, Jan. 2018.

[12] L. Ball, G. Ewan, and N. Coull, "Undermining: Social engineering using open source intelligence gathering," in *Proc. Int. Conf. Knowl. Discovery Inf. Retr.*, 2012, pp. 275–280.

[13] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," in *Proc. 13th AAAI Conf. Artif. Intell. (AAAI)*, 2016, pp. 2972–2978.

[14] J. Simola, "Privacy issues and critical infrastructure protection," in *Emerging Cyber Threats and Cognitive Vulnerabilities*, V. Benson and J. Mcalaney, Eds. Academic, 2020, pp. 197–226.

[15] M. Kandias, L. Mitrou, V. Stavrou, and D. Gritzalis, "Which side are you on? A new panopticon vs. privacy," in *Proc. IEEE Int. Conf. Secur. Cryptogr. (SECRYPT)*, Reykjavik, Iceland, Jul. 2013, pp. 1–13.

[16] L. R. Betts and K. A. Spenser, "Developing the cyber victimization experiences and cyberbullying behaviors scales," *J. Genet. Psychol.*, vol. 178, no. 3, pp. 147–164, May 2017.

[17] J. Pastor-Galindo, P. Nespoli, F. G. Mármol, and G. M. Pérez, "OSINT is the next Internet goldmine: Spain as an unexplored territory," in *Proc. 5th Nat. Conf. Cybersecur. (JNIC)*, Cáceres, Spain, 2019.

[18] F. Tabatabaei and D. Wells, "Osint in the context of cyber-security," in *Open Source Intelligence Investigation: From Strategy to Implementation*, B. Akhgar, P. S. Bayerl, and F. Sampson, Eds. Cham, Switzerland: Springer, 2016, pp. 213–231.

[19] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Quart.*, vol. 36, no. 4, pp. 1165–1188, 2012.

[20] V. Santarcangelo, G. Oddo, M. Pilato, F. Valenti, and C. Fornaro, "Social opinion mining: An approach for Italian language," in *Proc. 3rd Int. Conf. Future Internet Things Cloud*, Rome, Italy, Aug. 2015, pp. 693–697.

[21] M. Kandias, D. Gritzalis, V. Stavrou, and K. Nikoloulis, "Stress level detection via OSN usage pattern and chronicity analysis: An OSINT threat intelligence module," *Comput. Security*, vol. 69, pp. 3–17, Aug. 2017.

[22] B. Senekal and E. Kotzé, "Open source intelligence (OSINT) for conflict monitoring in contemporary South Africa: Challenges and opportunities in a big data context," *Afr. Secur. Rev.*, vol. 28, no. 1, pp. 19–37, Jan. 2019.

[23] D.-Y. Kao, Y.-T. Chao, F. Tsai, and C.-Y. Huang, "Digital evidence analytics applied in cybercrime investigations," in *Proc. IEEE Conf. Appl., Inf. Netw. Secur. (AINS)*, Nov. 2018, pp. 117–122.

[24] R. P. Pastor and H. L. Larsen, "Scanning of open data for detection of emerging organized crime threats—The ePOOLICE project," in *Using Open Data to Detect Organized Crime Threats*. Cham, Switzerland: Springer, 2017, pp. 47–71.

[25] C. Aliprandi, J. Arraiza Irujo, M. Cuadros, S. Maier, F. Melero, and M. Raffaelli, "Caper: Collaborative information, acquisition, processing, exploitation and reporting for the prevention of organised crime," in *HCI International 2014—Posters' Extended Abstracts*, C. Stephanidis, Ed. Cham, Switzerland: Springer, 2014, pp. 147–152.

[26] T. Delavallade, P. Bertrand, and V. Thouvenot, "Extracting future crime indicators from social media," in *Using Open Data to Detect Organized Crime Threats*. Cham, Switzerland: Springer, 2017, pp. 167–198.

[27] M. J. Hernández, C. C. Pinzón, D. O. Díaz, J. C. C. García, and R. A. Pinto, "Open source intelligence (OSINT) in a colombian context and sentiment analysys," *Rev. V'inculos, Ciencia, Tecnol. Sociedad*, vol. 15, no. 2, pp. 195–214, 2018.

[28] *Diversity Enhacements for Security Information and Event Management Project*. Accessed: Jan. 9, 2020. [Online]. Available: http://disiem-project.eu/

[29] S. Lee and T. Shon, "Open source intelligence base cyber threat inspection framework for critical infrastructures," in *Proc. Future Technol. Conf. (FTC)*, San Francisco, CA, USA, Dec. 2016, pp. 1030–1033.

[30] M. Edwards, R. Larson, B. Green, A. Rashid, and A. Baron, "Panning for gold: Automatically analysing online social engineering attack surfaces," *Comput. Security*, vol. 69, pp. 18–34, Aug. 2017.

[31] M. G. Lozano, J. Brynielsson, U. Franke, M. Rosell, E. Tjornhammar, S. Varga, and V. Vlassov, "Veracity assessment of online data," *Decis. Support Syst.*, vol. 129, Feb. 2020, Art. no. 113132.

[32] B. L. W. Wong, "Fluidity and rigour: Addressing the design considerations for osint tools and processes," in *Open Source Intelligence Investigation: From Strategy to Implementation*, B. Akhgar, P. S. Bayerl, and F. Sampson, Eds. Cham, Switzerland: Springer, 2016, pp. 167–185.

[33] G. Kalpakis, T. Tsikrika, N. Cunningham, C. Iliou, S. Vrochidis, J. Middleton, and I. Kompatsiaris, *OSINT and the Dark Web*. Cham, Switzerland: Springer, 2016, pp. 111–132.

[34] M. K. Bergman, "White Paper: The deep Web: Surfacing hidden value," *J. Electron. Publishing*, vol. 7, no. 1, Aug. 2001.

[35] M. Schafer, M. Fuchs, M. Strohmeier, M. Engel, M. Liechti, and V. Lenders, "BlackWidow: Monitoring the dark Web for cyber security information," in *Proc. 11th Int. Conf. Cyber Conflict (CyCon)*, Tallinn, Estonia, May 2019, pp. 1–21.

[36] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, Apr. 2015.

[37] A. Barnea, "Big data and counterintelligence in western countries," *Int. J. Intell. Counter Intell.*, vol. 32, no. 3, pp. 433–447, Jul. 2019.

[38] T. Day, H. Gibson, and S. Ramwell, "Fusion of OSINT and non-OSINT data," in *Open Source Intelligence Investigation*. Cham, Switzerland: Springer, 2016, pp. 133–152.

[39] C. S. Fleisher, "Using open source data in developing competitive and marketing intelligence," *Eur. J. Marketing*, vol. 42, no. 7/8, pp. 852–866, Jul. 2008.

[40] F. G. Marmol, M. G. Perez, and G. M. Perez, "Reporting offensive content in social networks: Toward a reputation-based assessment approach," *IEEE Internet Comput.*, vol. 18, no. 2, pp. 32–40, Mar. 2014.

[41] S. Gong, J. Cho, and C. Lee, "A reliability comparison method for OSINT validity analysis," *IEEE Trans. Ind. Informat.*, vol. 14, no. 12, pp. 5428–5435, Dec. 2018.

[42] M. Zago, P. Nespoli, D. Papamartzivanos, M. G. Perez, F. G. Marmol, G. Kambourakis, and G. M. Perez, "Screening out social bots interference: Are there any silver bullets?" *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 98–104, Aug. 2019.

[43] G. R. Weir, "The limitations of automating osint: Understanding the question, not the answer," in *Automating Open Source Intelligence*, R. Layton and P. A. Watters, Eds. Boston, MA, USA: Syngress, 2016, pp. 159–169.

[44] P. Casanovas, "Cyber warfare and organised crime. A regulatory model and meta-model for open source intelligence (OSINT)," in *Ethics and Policies for Cyber Operations*. Cham, Switzerland: Springer, 2017, pp. 139–167.

[45] H. Bean, "Is open source intelligence an ethical issue?" in *Research in Social Problems and Public Policy*, vol. 19, S. Maret, Ed. Bingley, U.K.: Emerald Group Publishing Limited, 2011, pp. 385–402.

[46] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*. Boston, MA, USA: Springer, 2012, pp. 415–463.

[47] P. Ranade, S. Mittal, A. Joshi, and K. Joshi, "Using deep neural networks to translate multi-lingual threat intelligence," in *Proc. IEEE Int. Conf. Intell. Secur. Inform. (ISI)*, Nov. 2018, pp. 238–243.

[48] Y. Ghazi, Z. Anwar, R. Mumtaz, S. Saleem, and A. Tahir, "A supervised machine learning based approach for automatically extracting high-level threat intelligence from unstructured sources," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, Dec. 2018, pp. 129–134.

[49] S. Noubours, A. Pritzkau, and U. Schade, "NLP as an essential ingredient of effective OSINT frameworks," in *Proc. Mil. Commun. Inf. Syst. Conf.*, Oct. 2013, pp. 1–7.

[50] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," *SIGKDD Explor. Newslett.*, vol. 17, no. 2, pp. 1–16, Feb. 2016.

[51] T. Vopham, J. E. Hart, F. Laden, and Y. Y. Chiang, "Emerging trends in geospatial artificial intelligence (geoAI): Potential applications for environmental epidemiology," *Environ. Health*, vol. 17, no. 1, Apr. 2018.

[52] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics—Challenges in topic discovery, data collection, and data preparation," *Int. J. Inf. Manage.*, vol. 39, pp. 156–168, Apr. 2018.

[53] L. Serrano, M. Bouzid, T. Charnois, S. Brunessaux, and B. Grilheres, "Events extraction and aggregation for open source intelligence: From text to knowledge," *Proc. Int. Conf. Tools Artif. Intell. (ICTAI)*, 2013, pp. 518–523.

[54] N. Kim, S. Lee, H. Cho, B.-I. Kim, and M. Jun, "Design of a cyber threat information collection system for cyber attack correlation," in *Proc. Int. Conf. Platform Technol. Service (PlatCon)*, Jan. 2018, pp. 1–6.

[55] S. Pournouri, S. Zargari, and B. Akhgar, "An investigation of using classification techniques in prediction of type of targets in cyber attacks," in *Proc. IEEE 12th Int. Conf. Global Secur., Saf. Sustainab. (ICGS3)*, Jan. 2019, pp. 202–212.

[56] I. Deliu, C. Leichter, and K. Franke, "Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 3648–3656.

[57] G. de la Torre-Abaitua, L. F. Lago-Fernández, and D. Arroyo, "A compression based framework for the detection of anomalies in heterogeneous data sources," 2019, *arXiv:1908.00417*. [Online]. Available: https://arxiv.org/abs/1908.00417

[58] R. Azevedo, I. Medeiros, and A. Bessani, "PURE: Generating quality threat intelligence by clustering and correlating OSINT," in *Proc. 18th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun./13th IEEE Int. Conf. Big Data Sci. Eng. (TrustCom/BigDataSE)*, Aug. 2019, pp. 483–490.

[59] M.-H. Wang, M.-H. Tsai, W.-C. Yang, and C.-L. Lei, "Infection categorization using deep autoencoder," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2018, pp. 1–2.

[60] H. Pellet, S. Shiaeles, and S. Stavrou, "Localising social network users and profiling their movement," *Comput. Secur.*, vol. 81, pp. 49–57, Mar. 2019.

[61] R. Wang, W. Ji, M. Liu, X. Wang, J. Weng, S. Deng, S. Gao, and C.-A. Yuan, "Review on mining data from multiple data sources," *Pattern Recognit. Lett.*, vol. 109, pp. 120–128, Jul. 2018.

[62] R. Layton, C. Perez, B. Birregah, P. Watters, and M. Lemercier, "Indirect information linkage for OSINT through authorship analysis of aliases," in *Trends and Applications in Knowledge Discovery and Data Mining*, J. Li, L. Cao, C. Wang, K. C. Tan, B. Liu, J. Pei, and V. S. Tseng, Eds. Berlin, Germany: Springer, 2013, pp. 36–46.

[63] A. Chaabane, P. Manils, and M. A. Kaafar, "Digging into anonymous traffic: A deep analysis of the Tor anonymizing network," in *Proc. 4th Int. Conf. Netw. Syst. Secur. (NSS)*, Sep. 2010, pp. 167–174.

[64] S. Pastrana, A. Hutchings, A. Caines, and P. Buttery, "Characterizing eve: Analysing cybercrime actors in a large underground forum," in *Research in Attacks, Intrusions, and Defenses*, M. Bailey, T. Holz, M. Stamatogiannakis, and S. Ioannidis, Eds. Cham, Switzerland: Springer, 2018, pp. 207–227.

[65] W. Tounsi and H. Rais, "A survey on technical threat intelligence in the age of sophisticated cyber attacks," *Comput. Secur.*, vol. 72, pp. 212–233, Jan. 2018.

[66] C. Sauerwein, I. Pekaric, M. Felderer, and R. Breu, "An analysis and classification of public information security data sources used in research and practice," *Comput. Secur.*, vol. 82, pp. 140–155, May 2019.

[67] R. Layton, "Relative cyberattack attribution," in *Automating Open Source Intelligence: Algorithms for OSINT*, R. Layton and P. A. Watters, Eds. Boston, MA, USA: Syngress, 2016, pp. 37–60.

[68] V. Mavroeidis and S. Bromander, "Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence," in *Proc. Eur. Intell. Secur. Inform. Conf. (EISIC)*, Athens, Greece, Sep. 2017, pp. 91–98.

[69] S. Bromander, A. Jøsang, and M. Eian, "Semantic cyberthreat modelling," in *Proc. 11th Conf. Semantic Technol. Intell., Defense, Secur.*, Fairfax, VA, USA, Nov. 2016, pp. 74–78.

[70] O. Akinrolabu, I. Agrafiotis, and A. Erola, "The challenge of detecting sophisticated attacks: Insights from SOC analysts," in *Proc. 13th Int. Conf. Availability, Rel. Secur. (ARES)*, 2018, pp. 55:1–55:9.

[71] D. Lande and E. Shnurko-Tabakova, "OSINT as a part of cyber defense system," *Theor. Appl. Cybersecur.*, vol. 1, no. 1, 2019.

[72] B. Akhgar, "Osint as an integral part of the national security apparatus," in *Open Source Intelligence Investigation: From Strategy to Implementation*, B. Akhgar, P. S. Bayerl, and F. Sampson, Eds. Cham, Switzerland: Springer, 2016, pp. 3–9.

[73] J. Chae, D. Graham, A. Henderson, M. Matthews, J. Orcutt, and M. S. Song, "A system approach for evaluating current and emerging army open-source intelligence tools," in *Proc. IEEE Int. Syst. Conf. (SysCon)*, Apr. 2019, pp. 1–5.

[74] D. Trottier, "Open source intelligence, social media and law enforcement: Visions, constraints and critiques," *Eur. J. Cultural Stud.*, vol. 18, nos. 4–5, pp. 530–547, Aug. 2015.

[75] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature Commun.*, vol. 10, no. 1, p. 3069, 2019.

[76] R. S. Portnoff, S. Afroz, G. Durrett, J. K. Kummerfeld, T. Berg-Kirkpatrick, D. Mccoy, K. Levchenko, and V. Paxson, "Tools for automated analysis of cybercriminal markets," in *Proc. 26th Int. Conf. World Wide Web (WWW)*, 2017, pp. 657–666.

[77] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, "Web data extraction, applications and techniques: A survey," *Knowl.-Based Syst.*, vol. 70, pp. 301–323, Nov. 2014.

[78] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2017.

[79] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, Apr. 2017.

[80] C. Eldridge, C. Hobbs, and M. Moran, "Fusing algorithms and analysts: Open-source intelligence in the age of 'big data,'" *Intell. Nat. Secur.*, vol. 33, no. 3, pp. 391–406, Apr. 2018.

[81] X. Yin, J. Han, and P. Yu, "Truth discovery with multiple conflicting information providers on the Web," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 6, pp. 796–808, Jun. 2008.

[82] A. S. Hulnick, "The dilemma of open sources intelligence: Is OSINT really intelligence?" in *The Oxford Handbook of National Security Intelligence*, L. K. Johnson, Ed. Oxford, U.K.: Oxford Univ. Press, Sep. 2010.

[83] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, Sep. 2017.

[84] B. H. Miller, "Open source intelligence (OSINT): An oxymoron?" *Int. J. Intell. Counter Intell.*, vol. 31, no. 4, pp. 702–719, Oct. 2018.

[85] G. Suarez-Tangil, M. Edwards, C. Peersman, G. Stringhini, A. Rashid, and M. Whitty, "Automatically dismantling online dating fraud," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1128–1137, 2020.

[86] J. Rajamäki and J. Simola, "How to apply privacy by design in osint and big data analytics?" in *Proc. Eur. Conf. Inf. Warfare Secur. (ECCWS)*, Jul. 2019, pp. 364–371.

[87] J. H. Hoepman, "Privacy design strategies," in *Proc. IFIP Adv. Inf. Commun. Technol.*, vol. 428, 2014, pp. 446–459.

[88] G. Hribar, I. Podbregar, and T. Ivanuša, "OSINT: A grey zone?" *Int. J. Intell. Counter Intell.*, vol. 27, no. 3, pp. 529–549, 2014.

[89] Q. Eijkman and D. Weggemans, "Open source intelligence and privacy dilemmas: Is it time to reassess state accountability?" *Secur. Hum. Rights*, vol. 23, no. 4, pp. 285–296, 2013.

[90] P. Mitzias, I. Kompatsiaris, E. Kontopoulos, J. Staite, T. Day, G. Kalpakis, T. Tsikrika, H. Gibson, S. Vrochidis, and B. Akhgar, "Deploying semantic Web technologies for information fusion of terrorism-related content and threat detection on the Web," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI) Companion*, 2019, pp. 193–199.

[91] G. W. Giumetti and R. M. Kowalski, "Cyberbullying matters: Examining the incremental impact of cyberbullying on outcomes over and above traditional bullying in North America," in *Cyberbullying Across the Globe*. Cham, Switzerland: Springer, 2016, pp. 117–130.

[92] E. M. Romera, M. Herrera-López, J. A. Casas, R. O. Ruiz, and R. Del Rey, "How much do adolescents cybergossip? Scale development and validation in Spain and Colombia," *Frontiers Psychol.*, vol. 9, pp. 1–10, Feb. 2018.

[93] L. Benes, "OSINT, new technologies, education: Expanding opportunities and threats. A new paradigm," *J. Strategic Secur.*, vol. 6, no. 3, pp. 22–37, Sep. 2013.

[94] A. López-Martínez, J. A. García-Díaz, R. Valencia-García, and A. Ruiz-Martínez, "CyberDect. A novel approach for cyberbullying detection on Twitter," in *Technologies and Innovation*. Cham, Switzerland: Springer, 2019, pp. 109–121.

JAVIER PASTOR-GALINDO received the B.Sc. and M.Sc. degrees in computer science from the University of Murcia, Spain. In 2019, he was granted with an FPU Predoctoral Contract by the Spanish Ministry of Science, Innovation and Universities, to develop his Ph.D. with the Department of Information and Communications Engineering, University of Murcia. His research interests focus on open source intelligence (OSINT), security, and privacy.

PANTALEONE NESPOLI received the B.Sc. and master's degrees in computer engineering from the University of Napoli Federico II, Italy. He is currently pursuing the Ph.D. degree with the University of Murcia, Spain. His research interests include information and communication systems security; more specifically network security, intrusion detection and response systems, and security information and event management.

FÉLIX GÓMEZ MÁRMOL received the M.Sc. and Ph.D. degrees in computer engineering from the University of Murcia, Spain. He is currently a Researcher with the Department of Information and Communications Engineering, University of Murcia. His research interests include cybersecurity, the Internet of Things, machine learning, and bio-inspired algorithms.

GREGORIO MARTÍNEZ PÉREZ is currently a Full Professor with the Department of Information and Communications Engineering, University of Murcia, Spain. His scientific activity is mainly devoted to cybersecurity, privacy, and networking. He is working on different national and European IST research projects (25 in the last decade) on these topics, being the principal investigator in most of them. He has published over 160 papers in national and international conference proceedings, magazines, and journals.