JOURNAL OF CYBERSECURITY

Journal of Cybersecurity, 2022, 1–15
https://doi.org/10.1093/cybsec/tyac003
Research paper

Research paper

Accessible from the open web: a qualitative analysis of the available open-source information involving cyber security and critical infrastructure

Yuxuan (Cicilia) Zhang*, Richard Frank, Noelle Warkentin and Naomi Zakimi

International CyberCrime Research Center (ICCRC), School of Criminology, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada

*Correspondence address. International CyberCrime Research Center (ICCRC), School of Criminology, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada, V5A 1S6; E-mail: ciciliaz@sfu.ca

Received 7 August 2021; revised 8 March 2022; accepted 18 March 2022

Abstract

In order to efficiently manage and operate industrial-level production, an increasing number of industrial devices and critical infrastructure (CI) are now connected to the internet, exposed to malicious hackers and cyberterrorists who aim to cause significant damage to institutions and countries. Throughout the various stages of a cyber-attack, Open-source Intelligence (OSINT) tools could gather data from various publicly available platforms, and thus help hackers identify vulnerabilities and develop malware and attack strategies against targeted CI sectors. The purpose of the current study is to explore and identify the types of OSINT data that are useful for malicious individuals intending to conduct cyber-attacks against the CI industry. Applying and searching keyword queries in four open-source surface web platforms (Google, YouTube, Reddit, and Shodan), search results published between 2015 and 2020 were reviewed and qualitatively analyzed to categorize CI information that could be useful to hackers. Over 4000 results were analyzed from the open-source websites, 250 of which were found to provide information related to hacking and/or cybersecurity of CI facilities to malicious actors. Using thematic content analysis, we identified three major types of data malicious attackers could retrieve using OSINT tools: indirect reconnaissance data, proof-of-concept codes, and educational materials. The thematic results from this study reveal an increasing amount of open-source information useful for malicious attackers against industrial devices, as well as the need for programs, training, and policies required to protect and secure industrial systems and CI.

Key words: open-source intelligence, critical Infrastructure, cybersecurity, cyber-attack, industrial systems, qualitative research

Introduction

Over the past decade, the development of automated decision-making and remote accessing and controlling of industrial devices has benefitted the manufacturers in various aspects; these remote-control systems have reduced manual labor and increased the overall efficiency of production [1–4]. Valuing convenience brought by technology, countries like Canada are now connecting every major industry including critical infrastructure (CI) to the cyber-world [5, 6]. It is no longer rare to find CI and industrial facilities such as Industrial Control Systems (ICS)[1] or Supervisory Control and Data Acquisition

1 Control systems including supervisory control and data acquisition systems and other smaller systems or controllers used in industrial production or CI sectors.

(SCADA) systems transitioning from an air-gapped network environment to externally connected with internet networks, yielding more opportunities for malicious attacks and cyber warfare [7–9].

Within the current ICS system, multiple devices including, but not limited to, the programmable logic controllers (PLCs), remote terminal units (RTUs), master terminal units (MTUs), three-term controllers (PID controllers), and SCADA servers, all require internet to complete the automation process [3, 4]. Specific communication protocols for industrial production are also coded, such as the Modbus protocol, Transmission Control Protocol/Internet Protocol (TCP/IP), Distributed Network Protocol (DNP3), and Common Industrial Protocol (CIP) [10]. Functioning interdependently in the system, a successful intrusion in any of these devices or protocols would compromise the entire system and ultimately result in cascading failure of multiple interconnected SCADA systems in the industrial sector. As systems start to fail after successful attacks, broad-scale impacts may arise, including financial and physical damage to the factory, loss of lives, and damage to the reputation and market share of the companies [14].

Part of what makes the CI devices at higher risk of security breaches is the availability of Open-source Intelligence (OSINT) techniques to gather information relevant to CI systems and plan out attacks. The use of OSINT during data gathering by both hackers and security professionals typically involves the use of publicly available sources such as online articles, search engines, social-networking platforms, video recordings, as well as personal blogs [11]. With widely available information related, but not restricted, to cybersecurity and CI, such data disclosure could provide hackers opportunities to gather and learn from these online data and pinpoint desirable targets and strategies for successful cyberattacks [12, 13]. Although previous literature has pointed out some areas of concern with regards to the danger of OSINT data and cyber security of industrial systems in the CI sector, studies about the types of CI-related information one can gather from open-source surface web platforms such as Google, YouTube, Shodan, and journals or conferences have rarely attracted researchers' attention. Serving as one of the main platforms for social media sites and information sharing, these websites may contain fruitful resources, tools, and data useful for malicious hackers wishing to conduct cyber–physical attacks.

This paper aims to explore the major types and potential use of data obtainable by malicious individuals targeting the CI industry from the openly accessible surface web. We used publicly available OSINT resources to retrieve and analyze data providing insight about cyber attacks against CI systems, such as step-by-step tutorials, technical analysis of zero-day exploits, or information about vulnerable devices, exploits, and open-source tools. The findings of the current study may provide insight on the potential threat surface web data could pose to CI facilities, as well as aid in the development of more rigorous mitigation strategies and recommendations to CI vendors to prevent from future cyber attacks.

In the following sections, we first present a brief overview of the current literature concerning OSINT and cybersecurity of CI. Then, we discuss the methods and data used followed by the findings, which are divided into three main themes: indirect reconnaissance data, proof-of-concept (PoC) codes, and educational materials. Last, we conclude by discussing the implications and limitations of the study.

Related Work

The cyber attack kill-chain

Extensive research has been carried out to understand the process and purpose of various cyber attacks [1, 2, 15, 16]. Building upon the increasing interest in cybersecurity and cyber attacks, Hutchins *et al*. [16] proposed a kill chain model with seven end-to-end stages of a cyber attack, informing a starting point for research layering out what and how hackers need to do prior to a successful cyber attack. Upon preparation of a cyber attack, hackers will often follow this path: reconnaissance, weaponization, delivery, exploitation, installation, command and control, and finally to act on an objective [16].

The first stage in the kill-chain, reconnaissance, refers to the phase when hackers gather information about the target in order to plan out the attack [1, 2, 15]. Reconnaissance data are vital to a successful attack; attackers are required to obtain information related to communication configurations, ports and exploitable vulnerabilities, as well as ideal devices granting access to the system in order to proceed to later attack stages [1].

Using data gathered in the research stage, hackers then develop custom malware against the desirable target and seek suitable means to distribute the exploit in the weaponization (stage 2) and delivery[2] phase [16]. The successful delivery of the exploit would trigger the exploitation stage (stage 4), where malicious codes start running and infecting the host system, followed by the installation of malware (stage 5) granting access to the hackers [16]. The intrusion of malware into the host provides hackers with full command and control (stage 6) over the target system, allowing them to take actions in order to achieve their goals. In attacks targeting CI and cyber–physical systems such as power grids and industrial manufacturing plants, hackers would intend to achieve physical objectives (stage 7) such as disruption of service and/or destruction of the facility by commanding and controlling the devices [15].

OSINT network scanners and data gathering

Advances in the Internet of Things (IoT) and exposure of industrial devices to the internet is now putting systems in the CI at higher levels of risk. Particularly, the issue became problematic when an increasing amount of ICS-related open-source information and tools are available to the public. Researchers have begun to see that data gathered through public domain could help both cybersecurity professionals and malicious hackers significantly in the reconnaissance stage [17, 18]. To demonstrate, Samtani *et al*. [17] conducted a study focusing on gathering hacking-related data from hacker communities. The findings suggested vast amounts of information related to hacking and cyber-attacks could be obtained by malicious individuals from online sources [17].

Research focused on the impact of open-source tools such as network scanners on cybersecurity has also been studied extensively [4, 19, 20]. Publicly available network scanners, with their ability to search for internet-connected industrial devices, e.g. can be exploited by malicious attackers [19]. According to Bodeheim *et al*. [19], in fact, Shodan was able to collect information on ICS devices through establishing communications with open service ports. Upon successful communication between the Shodan signal and the service port, the search engine would record information including the device's location, IP address, as well as detailed data such as open ports, services, and protocols exploitable by hackers. This rendered Shodan, as one of the most useful network scanners available to the public, identifying all of the four PLC devices and the static IP addresses assigned to each of the controller, demonstrating the search engine's ability to discover internet-connected industrial devices [19].

Recent studies of Shodan have further confirmed Bodeheim *et al*.'s result [4, 20]. In 2018, researchers found that >500 000 internet-

2 According to Hutchins *et al*., the delivery phase is the third stage of a cyber attack kill chain [14].

connected SCADA devices are discoverable through Shodan; numerous devices were detected to have existing exploitable features, such as the use of default credentials and unpatched system vulnerabilities [4]. Another study conducted by Chen *et al.* [20] discovered Shodan's capability of identifying and indexing all six honeypots they released to the public, demonstrating the exploitable scanning abilities and functions featured in OSINT tools during reconnaissance stage.

OSINT and social engineering

Apart from focusing on the data-gathering process using OSINT tools, previous research has also focused on social engineering-based cyber attacks and human-centric factors affecting ICS [21, 22]. Social engineering is referred to as a "form of deception in which an attacker attempts to fraudulently acquire sensitive information from a victim by impersonating a trustworthy entity" [21]. For malicious attackers, employees with access and control over internal networks and CI facilities are often targets to gather information and perform attacks on. Through leveraging different OSINT tools, social networking information and techniques, malicious offenders can target vulnerable employees working in the CI sectors and exploit their trust and personal information for cyber attacks [22–24]. In some scenarios, malicious attackers would conduct social engineering through forms of coercion, intimidation, or blackmailing the targets using the information they obtained online [2].

The selection of ideal targets among the list of employees working in the CI sector is not random; only employees with select features or characteristics are considered as ideal targets [18, 23–25]. Particularly, hackers would look for the ranking or position of the employee within the company. Key employees working in target industrial companies would attract most of the attackers' attention as they often possess confidential information and administrator usernames and passwords that would grant privileged access to control systems [18].

Other demographic characteristics such as gender, cyber awareness and skills, or frequency and amount of personal information employees shared online are also important features to consider [23–25]. For example, Jagatic *et al.* [21] conducted an experiment sending out phishing emails to a group of targeted participants; the findings suggested female employees and individuals who do not have technology backgrounds are more likely to become victims of social engineering. Through manual or automated data gathering methods, attackers could easily identify employees with desirable traits, which could signal their low resistance against social engineering and, therefore, deploy phishing emails to gain trust from these victims [23]. The information gathered from the public domain containing organization's contact information, identification of an individual as an affiliate or employee in the company, as well as social media connections between different colleagues are also rich sources of information for malicious attackers, allowing them to curate scenarios less likely to be detected by targeted victims as fraudulent [25].

Methodology

Research question

Resolving the lack of research in the types of open-source data related to CI searchable from surface web platforms was the aim of the current paper. Previous studies related to the use of OSINT data have principally focused on developing and testing data mining or deep-learning models used to detect and track communications and motivations shared by criminals online [13]. Although studies concerning OSINT data-gathering for the purpose of malicious cyberattacks are on the rise, most research is concentrated on exploring personal profiles useful for social engineering attacks, or illegal open-source materials shared by cybercriminals on the Dark Web [13, 26].

To our knowledge, there has been a lack of research focusing on OSINT data related to the cybersecurity of CI facilities circulating in the surface web. While studies have discussed OSINT tools, techniques, and the benefit of them in areas of risk assessment or cyber–forensic analysis, the majority of tools were examined and applied in platforms including (but not limited to) Twitter, Facebook, Dark Web, or Shodan [13, 19, 20, 26]. Indeed, current OSINT-related research rarely paid attention to the types of CI-related data offered in surface web search engines that could greatly benefit ill-intended hackers. As we have observed a substantive increase in the quantity and frequency of cyberattacks as well as hacking against organizations and businesses, the types of exploitable OSINT data retrieved from the surface web may pose a serious security threat on CI or other industrial devices. As such, we aimed to answer the following research questions:

RQ1. What types of CI-related data can be found from the surface web?

RQ2. How can these data be useful for malicious cyber-attackers?

Data sources and collection

The dataset for this research was collected based on results returned from keyword queries in four open-source search engines and websites including Google, Reddit, YouTube, and Shodan on the surface web. These four platforms were specifically selected for the study as they are all widely known public search platforms accessible for internet users to find information and have their questions answered. The platforms are also bounded by policies and legislations where sensitive information will be monitored and removed [27–29]. Open-source data from privacy-oriented programs such as the Darknet and DuckDuckGo were not investigated or included in the current study, as more effort and knowledge related to the programs are required in order to access contents hidden in the dark web. Privacy-oriented platforms are not monitored by governmental institutions or platform staff due to their heightened encryption and privacy settings, which they provide more opportunities for malicious individuals to openly share illegal hacking contents [30, 31]. Thus, those types of websites were excluded.

A purposive sampling technique was employed; a list of keyword queries relating to ICS and CI was composed and entered for the research. The set of keywords were discussed and reviewed by the authors, to identify any relevant queries related to the current research question; the keyword searches were conducted and reviewed by the researchers. Table 1 presents all the keywords bulk-searched from the targeted open-source websites in the current study. The default search setting was applied to all websites (e.g. Google, YouTube, and Reddit); no filters or advance search settings were applied. All posts were by relevance. The keyword search yielded over 4000 displayed results captured by Google[3], YouTube, and Reddit.

From among those results, criterion sampling was employed to filter out irrelevant webpages that were not of interest for this study. Only content published between 1 January 2015 and 15 July 2020 were recorded for further analysis, ensuring the information obtained online was up-to-date. To be included, results also had to provide

3 Only a certain number of most relevant results were displayed by Google keyword search, with majority of the repetitive and non-relevant webpages were omitted.

Table 1: List of keyword queries searched using OSINT

Keyword set #1		Keyword set #2
(SCADA OR supervisory control and data acquisition) (Programmable logic controller OR programmable logic controllers OR PLC) (PID OR PID controller OR three-term controller) (RTU OR remote terminal unit) (Modbus OR DNP3) (Modicon OR Unitronics) (Eaton OR Eaton industrial OR Honeywell OR Midas gas detector) (CirCarLife OR Advantech OR Laquis) (SINEMA Siemens OR industrial OR server) (PROFIBUS OR Honeywell HART OR Simatic OR Schneider OR Cisco) (Infrastructure OR chemical OR dam OR emergency OR nuclear OR transportation OR water OR plant OR energy OR blackout OR electricity OR power OR gas OR industrial OR manufacturing cascading failure) (Industrial control system OR ICS OR CI) (PCS OR process control system OR advanced process control OR distributed control system OR distributed control systems OR DCS) (GE Automation OR OMRON industrial controller OR OMRON PLC OR Mitsubishi electric PLC) (Very small aperture terminal OR VSAT OR power grid OR smart grid) (Dragonfly OR Havex OR Industroyer OR Crashoverride OR Stuxnet OR Duqu OR BlackEnergy OR Triton OR Trisis OR EKANS OR MegaCortex)	**AND**	(Exploit OR vulnerability OR hack OR malware OR attack OR zero-day OR 0day OR access OR intrude)
		Total keywords (*N*) = 81

enough information for a malicious hacker about CI systems and information related to hacking these systems. For example, content reporting occurrences of cyberattacks will be excluded as it does not contain much information related to hacking techniques. On the other side, information that may have contained exploitable CI facilities, or technical analysis and in-depth discussion relating to CI mechanisms or security will be included for further analysis. Lastly, to be included in the study, contents also must be in English. Search results were reviewed by the authors.

Our initial plan was to unify the data collection method and evaluate the keyword search results in all of the publicly available search engines including Shodan. Operating differently than websites such as Google, Shodan only provided results when keyword queries contain specific models, protocols, or vendors in the CI sector. Complimenting on the network scanning nature of Shodan, a modified list of keywords including industrial devices' names, models, manufacturers, or numbers (e.g. Siemens S7, PLC, Modbus, SINEMA, and so on) commonly used in the current CI sectors was used for Shodan searches. Providing information on searchable CI facilities, the results related to CI devices was included in the sample. Each search feedback on Shodan was counted as one since the nature of the results are the same (e.g. 396 publicly searchable Modbus devices across the world) for each keyword; the keyword search yielded 28 displayed results from Shodan.

Data analysis

All relevant results displayed by search engines (Google and Shodan) were analyzed during the analysis phase ($n = 3530$). A random selection of 70 YouTube videos, as well as ~400 posts captured in Reddit were also reviewed[4]. Data that met the inclusion criteria were recorded in a spreadsheet for detailed coding and analysis. First, we conducted a title review; data that met the inclusion criteria at this stage were recorded in a spreadsheet for further analysis. Then, a first round of coding was conducted and any results that did not meet inclusion criteria were excluded. Coding and analysis of all results that emerged from the keyword queries were complete by the authors. The final sample included in this consisted of 228 webpages discoverable from Google and Shodan, 16 videos from YouTube, as well as one relevant Subreddit[5] community and five threads (Fig. 1). The codes and themes were developed and discussion of the results were hold between the authors during the progress. All of the contents listed in the final sample were analyzed and coded individually in the same spreadsheet, with the authors' names replaced with letters (e.g. *Author A*, *Author B*, *Website C*, and so on).

For the purpose of this study, a thematic content analysis was conducted to uncover the common types of information obtainable from OSINT resources by malicious hackers [32]. A total of four rounds of inductive coding were conducted to allow the development of descriptive codes. The results and the themes that emerged from the thematic analysis were presented and discussed with the group of authors to ensure consensus was reached. A total of six codes emerged from the data set and were later collapsed into three main themes.

Results

In this section, we detailed the types of data malicious attackers could retrieve from open-source web searches. A total of three main themes were identified after three rounds of coding: "Indirect reconnaissance data," "Proof-of-concept codes," and "Educational materials." Within the themes, three sub-themes were identified in "Indirect reconnaissance data," and two subthemes were categorized under the theme "Educational materials." A complete list of the themes, codes, and the frequency count emerged from the current data set can be found in Table 2 below. Our findings were consistent with results discussed in previous literature [1, 4, 15, 18, 33–35]. Although the

4 The random selection of 70 YouTube videos and approximately the same quantity of Reddit results was reviewed as both platforms' results were designed not to provide the total quantity returned from the search, making researchers unable to obtain accurate number of results.

5 Subreddit is referred to an online community residing within Reddit.

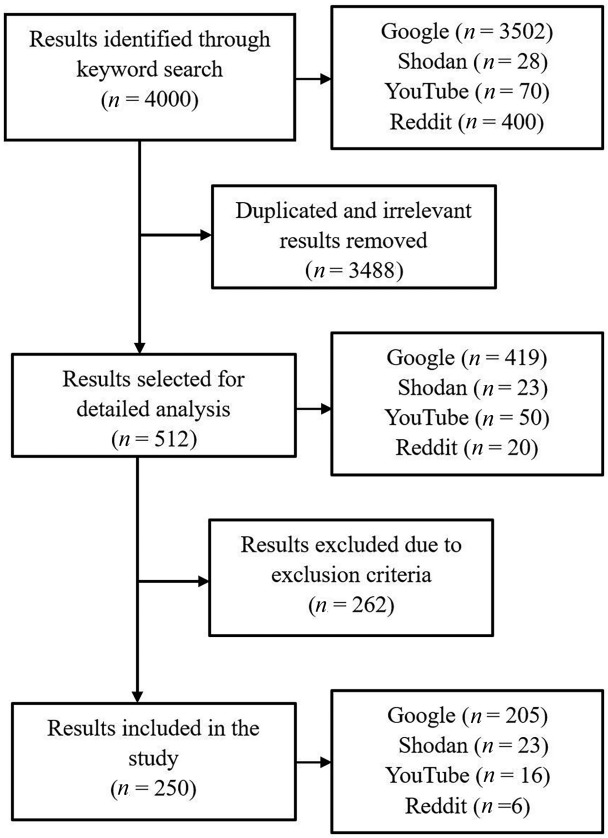


Figure 1: OSINT data results selection process

majority of the information from open-source websites was intended for cyber-security personnel or ethical hackers, the same set of data, such as threat-related information, hacking trainings, as well as published PoC exploits, can be used maliciously by hackers against CIs and vendors.

Indirect reconnaissance data

The majority of the information (70.4%) we found through OSINT search of keywords were data helping hackers indirectly during the research stage of cyber attack kill-chain ($n = 176$). The availability of such knowledge allows attackers to determine the appropriate hacking strategy, methods to avoid detection, the malware of choice, as well as ideal targets prior to the attack. A total of three forms of data were identified in the open web: (1) threat-related information; (2) hacking and reconnaissance tools; and (3) demonstration videos.

Threat-related information

General information about CI, vulnerable industrial devices, and potentially useful malware was the most prominent type of data one can look up in the public domain. Online communities such as Reddit allowed their users to share hacking-related information potentially related to CI. In one of the posts, *User A* shared a book with an updated version of programming code for members interested in learning coding and understanding cybersecurity. *User A* suggested in his post, that "some of the contents of the book cover how to program port scanners, reverse shells, your own botnet command and control center, extract EXIF information from image files, instantiate an anonymous browser in Python, and more." The content shared by users like *A* suggested that individuals could gather useful cyber security information from social networking communities.

Although Reddit allowed its members to ask questions with regards to existing bugs and exploitation techniques with each other, certain rules had to be followed and were enforced by volunteer moderators. The public nature of social media often required various platforms to strictly obey the rules and regulations enforced by the government. One subreddit was particularly concerned about the contents posted in the community and stated:

> *Avoid self-incriminating posts. Sometimes you might do some research that is ethically (and legally) questionable. Soliciting others to incriminate falls under this umbrella, as you would become co-conspirators. This is Reddit, and this is public. Use your brain. … No "Please hack X" posts. Save that shit for hack forums.* (*Subreddit R*, 2020)

Information released on governmental websites or journal and conference publications could provide insight for malicious attackers as well. For example, one *advisory B* provided both descriptions of the vulnerability and its affected device models and software versions: "A Heap-based Buffer Overflow was found in Emerson OpenEnterprise SCADA Server 2.83 (if Modbus or ROC Interfaces have been installed and are use) and all versions of OpenEnterprise 3.1–3.3.3, where a specially crafted script could execute code on the OpenEnterprise Server." Most of these sources would be beneficial to hackers in the planning phase when easy-to-target models and services, potential attack strategies, and available vulnerability for malware-development need to be decided.

Similar to the information posted in public advisories and articles, technical analysis reports were also capable of providing hackers vague instructions on how certain strategies and cyberattacks were employed on industrial devices. For instance, a *report C* analyzed the malware CrashOverride and discussed the features and registers of the backdoor module found in the exploit's artifact,

> *… reviewing memory during execution and analysis of other modules in the malware indicates that \Sessions\1\Windows\ appears multiple times, indicating that a check may be performed. The backdoor writes a file to either C:\Users\Public\ or C:\Users\<Executing User > .* (*Report C*, 2017)

This type of analysis was common in reports produced by technical analysts since security personnel were required to understand zero-day exploits prior to the development of patches and mitigation strategies for these vulnerabilities. Oftentimes, academic researchers were also interested in publishing journals or blogs about exploitable industrial devices, attempting to attract vendors' attention and thus implement mitigation strategies accordingly.

> *Another diagnostic command attacker can use is Read Device Identification as an attempt to gather information on Modbus device: A MODBUS request packed with function code 43 Read Device Identification will cause a MODBUS server to return the vendor name, product name, and version number. Additional information may also be provided in optional fields. An attacker sends the MODBUS request packet with function code 43 to all systems in the network and gathers intelligence that may be helpful in future attacks.* (*Researcher D*, 2019)

As described by *researcher D*, a hacker could obtain detailed information of a target device by injecting certain command toward the Modbus system.

Some public sources were able to give directions to hackers on social engineering techniques toward employees working in the target

Table 2: Primary codes, categories, and frequencies

Theme	Sub-themes	Frequency (%)
Indirect reconnaissance data	—Threat-related information	132 (52.8%)
	—hacking and reconnaissance tools	32 (12.8%)
	—demonstration videos	12 (4.8%)
PoC codes		40 (16.0%)
Educational materials	—"How to" tutorials	26 (10.4%)
	—training courses	8 (3.2%)
	Total:	250 (100%)

industrial company. Providing instructions on what pirates should do prior to hacking into SCADA systems in the cargo ships, *Presenter E* said:

> *…we are gonna pivot from the ship tracker sites into myship.com. … You can go in, look up any ship you want, find out who the crew are, find out who their ship mates are, and then social engineer the hell out of it. … And then pivot from there with whatever it is you do …* (*Presenter E*, 2018)

This kind of social engineering method was typically used in cyber–physical attacks; attackers could either gain the trust of the target employee and obtain their credentials to access into the industrial system, or they could gather the victims' private information and coerce the employee to become an insider and conspire the cyber attack together.

Other hackers could attempt to gain access to the industrial systems by using default credentials retrieved online. Default usernames and passwords were primarily found in publicly accessible user manuals of industrial devices: "After installation, log in with the user name 'admin' and the password 'admin'."[6] Compiled list of default credentials of industrial devices, such as the spreadsheet shared by *User G* in Fig. 2, can also be found. This information can pose risk to institutions still using default usernames and passwords for commanding and controlling their ICSs.

Hacking and reconnaissance tools

Not surprisingly, educational tools and software programs designed for ethical hackers were commonly shared on open-source websites. Simulation software of industrial devices were able to provide individuals opportunities to understand the algorithms and operational commands ensuring the functioning of the systems. One *website H* highlighted the educational version of a ladder logic[7] programming software used to operate industrial systems:

> *… [a] complete (not crippled version) software package for learning about PLC programming and for users to evaluate the power of Ladder Logic or Ladder + BASIC software programming. … The program files you created using the Educational version are identical to that of the Production version so you can write and test your entire program to make sure that it can do what you want …* (*Website H*, 2020)

According to this description, the PLC programs written in this free software are identical and applicable to PLC systems in the real world.

Other tools such as OSINT reconnaissance programs or penetration testing software provided to professional cybersecurity researchers could be downloaded by malicious attackers. *User I* shared a list of useful resources professional hackers could use when conducting vulnerability assessment, including tools capable of "brute-force the password used by S7[8] instances from a PCAP using a dictionary."

Research on the open-source scanner Shodan was able to provide reconnaissance information of internet-connected industrial devices. Results from Shodan were capable of giving general information such as the quantity of industrial devices and controllers identified on the internet, the locations of the devices, as well as top service manufacturers of the device. Figure 3 displays the search result for Allan-Bradley devices discovered by Shodan.

Detailed data on these devices are available for further inspection on Shodan. Figure 4 illustrates Shodan's ability to identify specific industrial devices connected to the internet. These types of information are helpful for both professional researchers and cyberterrorists in the early stages of attack in order to determine the ideal target and attack strategies.

The textual details of this specific logic controller displayed in Fig. 4 provided information about the IP address and the organization owning the device, open ports on the device, as well as services operating on each of the ports. Shodan also compiled a list of reported vulnerabilities one could potentially exploit against the device if the issues were unpatched by its vendor.

Demonstration videos

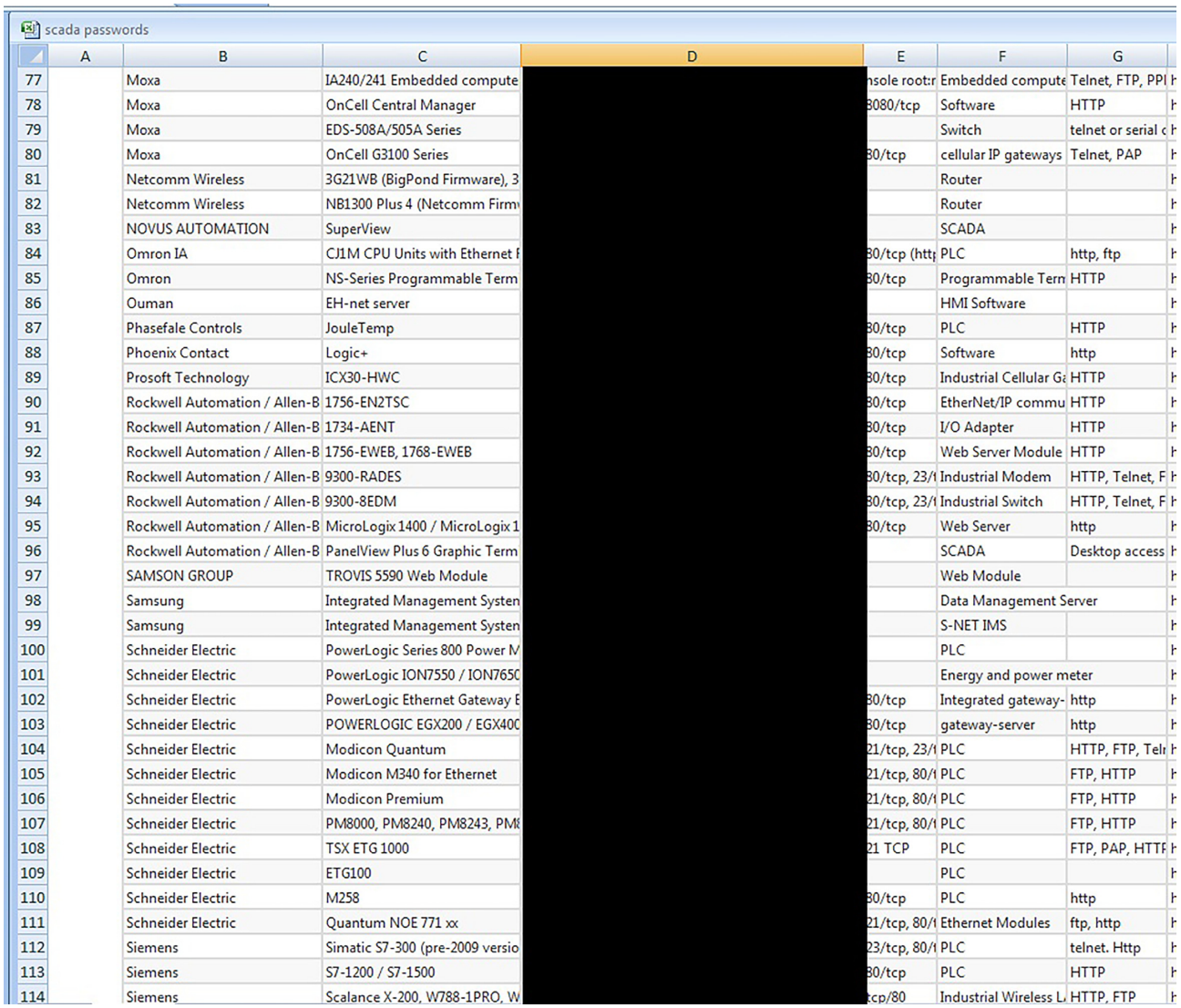
Videos demonstrating successful intrusion into industrial systems are uploaded by cybersecurity companies or individuals interested in hacking and computer programming. Among all of the uploaded videos, users often did not provide any verbal explanations to either the hacking process or the exploit; rather, these videos were mostly silent with snippets of the coding inputs and outputs presented throughout. Although brief descriptions of the content and types of the attack against industrial devices were provided, most of the users did not provide further information with regards to where and how to obtain a copy of the zero-day exploit or source codes of the malware. As shown in Fig. 5, *User J* demonstrated a successful multiple DoS attack against a virtual PLC model using Python in their YouTube channel *Claes*.

Cyber attacks targeting specific industrial devices or software were also performed by security labs and uploaded on YouTube or Vimeo as proof of vulnerability assessment results for their clients or cohorts. To illustrate, *Lab K* from ExCraft uploaded a video demon-

6 *User manual F*.

7 Ladder logic is referred to the programming language commonly used to program and control Programmable Logic Controllers (PLCs). For more information, see [36].

8 S7 is referred to the Siemens S7 PLC product series.



scada passwords

	A	B	C	D	E	F	G
77		Moxa	IA240/241 Embedded compute	[redacted]	nsole root:r	Embedded compute	Telnet, FTP, PPI
78		Moxa	OnCell Central Manager		8080/tcp	Software	HTTP
79		Moxa	EDS-508A/505A Series			Switch	telnet or serial c
80		Moxa	OnCell G3100 Series		80/tcp	cellular IP gateways	Telnet, PAP
81		Netcomm Wireless	3G21WB (BigPond Firmware), 3			Router	
82		Netcomm Wireless	NB1300 Plus 4 (Netcomm Firm			Router	
83		NOVUS AUTOMATION	SuperView			SCADA	
84		Omron IA	CJ1M CPU Units with Ethernet I		80/tcp (htt	PLC	http, ftp
85		Omron	NS-Series Programmable Term		80/tcp	Programmable Tern	HTTP
86		Ouman	EH-net server			HMI Software	
87		Phasefale Controls	JouleTemp		80/tcp	PLC	HTTP
88		Phoenix Contact	Logic+		80/tcp	Software	http
89		Prosoft Technology	ICX30-HWC		80/tcp	Industrial Cellular Ga	HTTP
90		Rockwell Automation / Allen-B	1756-EN2TSC		80/tcp	EtherNet/IP commu	HTTP
91		Rockwell Automation / Allen-B	1734-AENT		80/tcp	I/O Adapter	HTTP
92		Rockwell Automation / Allen-B	1756-EWEB, 1768-EWEB		80/tcp	Web Server Module	HTTP
93		Rockwell Automation / Allen-B	9300-RADES		80/tcp, 23/t	Industrial Modem	HTTP, Telnet, F
94		Rockwell Automation / Allen-B	9300-8EDM		80/tcp, 23/t	Industrial Switch	HTTP, Telnet, F
95		Rockwell Automation / Allen-B	MicroLogix 1400 / MicroLogix 1		80/tcp	Web Server	http
96		Rockwell Automation / Allen-B	PanelView Plus 6 Graphic Term			SCADA	Desktop access
97		SAMSON GROUP	TROVIS 5590 Web Module			Web Module	
98		Samsung	Integrated Management Systen			Data Management Server	
99		Samsung	Integrated Management Systen			S-NET IMS	
100		Schneider Electric	PowerLogic Series 800 Power M			PLC	
101		Schneider Electric	PowerLogic ION7550 / ION765C			Energy and power meter	
102		Schneider Electric	PowerLogic Ethernet Gateway E		80/tcp	Integrated gateway-	http
103		Schneider Electric	POWERLOGIC EGX200 / EGX40C		80/tcp	gateway-server	http
104		Schneider Electric	Modicon Quantum		21/tcp, 23/t	PLC	HTTP, FTP, Tel
105		Schneider Electric	Modicon M340 for Ethernet		21/tcp, 80/t	PLC	FTP, HTTP
106		Schneider Electric	Modicon Premium		21/tcp, 80/t	PLC	FTP, HTTP
107		Schneider Electric	PM8000, PM8240, PM8243, PM8		21/tcp, 80/t	PLC	FTP, HTTP
108		Schneider Electric	TSX ETG 1000		21 TCP	PLC	FTP, PAP, HTTP
109		Schneider Electric	ETG100			PLC	
110		Schneider Electric	M258		80/tcp	PLC	http
111		Schneider Electric	Quantum NOE 771 xx		21/tcp, 80/t	Ethernet Modules	ftp, http
112		Siemens	Simatic S7-300 (pre-2009 versio		23/tcp, 80/t	PLC	telnet. Http
113		Siemens	S7-1200 / S7-1500		80/tcp	PLC	HTTP
114		Siemens	Scalance X-200, W788-1PRO, W		tcp/80	Industrial Wireless L	HTTP, FTP

Figure 2: Partial list of default credentials shared by *User G* from personal Blog (Hackers-arise)

strating how Advantech WebAccess[9] Version 8.3.2 can be hacked and remotely controlled by hackers (Fig. 6).

PoC codes

The second major type of data retrievable from surface web platforms are PoC codes. The idea of sharing these PoC codes can be traced back to some early projects proposed by cybersecurity personnel in the past decade. Most significantly, Project Basecamp, presented in a cybersecurity conference in 2012 has proposed the idea of sharing PoC codes so the industry would become aware of the potential vulnerabilities in the field devices [37]. In the presentation, the event lead Peterson explained the team's motive for sharing the PoC codes:

> *… Eric Butler a few years later came up with this Firefox plugin called Firesheep now made it possible for anyone sitting in a coffee shop who could use a browser to hijack a session and guess what, it got people's attention. Even though everyone knew before, now that everyone could do it. Things change very quickly and those vendors that had done nothing about it very quickly added the capability to … solve this problem. … maybe we need a firesheep moment in PLC security.* [38]

The initial goal for sharing PoC codes was to allow security researchers to perform and understand the exploits, and eventually develop and enforce better mitigation strategies to protect industrial devices. With this intention, the PoC exploits can often be discovered in websites such as Exploit Database (EDB), Github, or attached within official advisories published by security companies or government websites.

For example, *User L* released a version of buffer overflow exploit on EDB, noting that the successful launch of the code would crash the Modbus slave PLCs. The details of the affected software, the tested version of the device, as well as the operating system used to perform the exploit were provided. *User L* also included a brief instruction on the steps required to run and reproduce the exploit. An excerpt of the PoC exploit is demonstrated in Fig. 7.

9 The Advantech WebAccess software allows industrial companies to connect the devices with internet and provides remote access to the ICS. See https://www.advantech.com/industrial-automation/webaccess for more details.

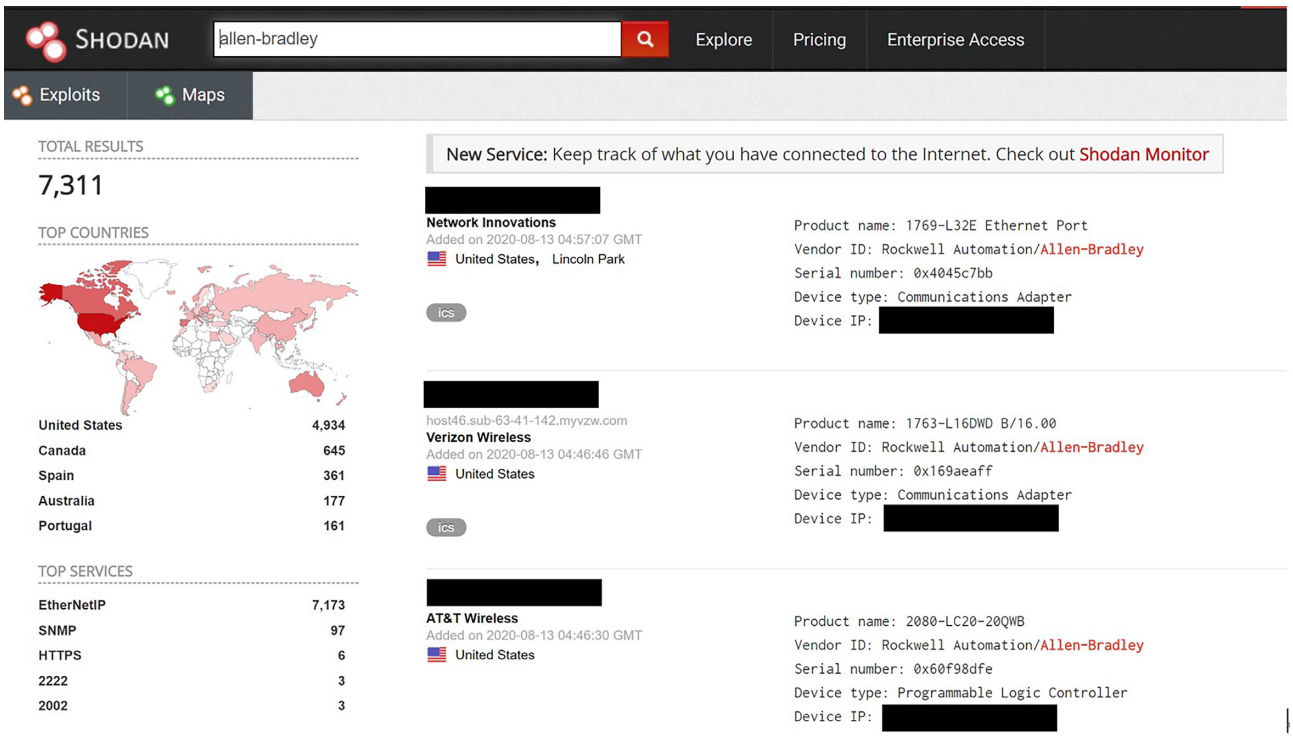


Figure 3: Shodan search result for Allan-Bradley devices (Shodan)

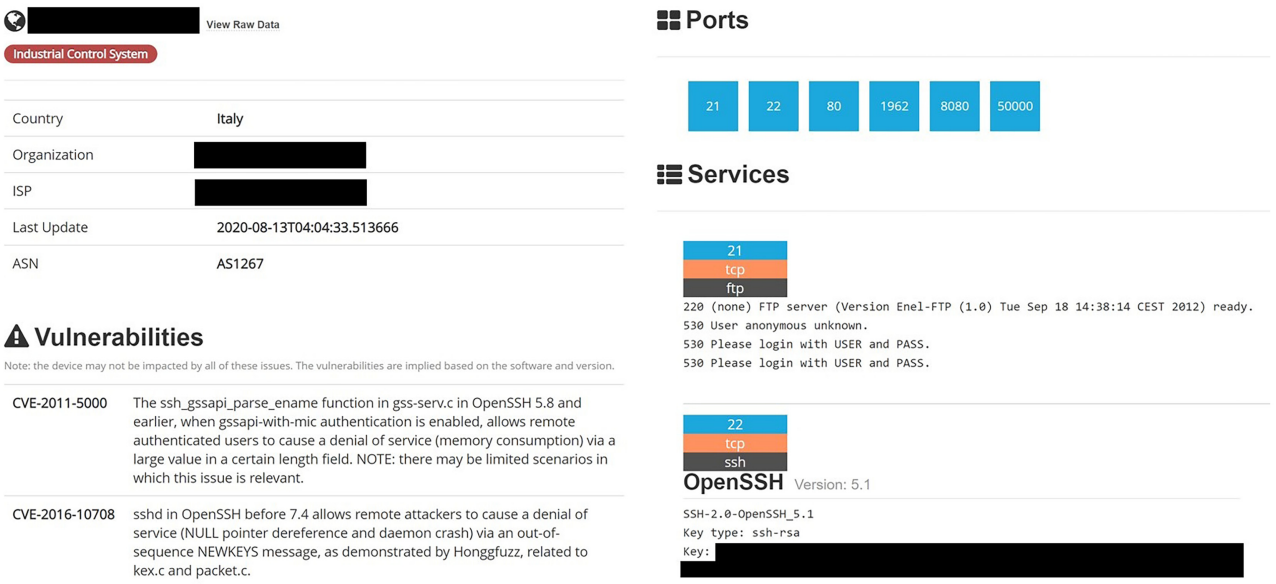


Figure 4: Detailed information of an industrial PLC device in Italy (Shodan)

Critical exploits of devices commonly installed in the industry could also be discovered in vulnerability reports. Figure 8 below illustrates a partial PoC exploit capable of inducing a denial-of-service attack on Schneider Electric's Modicon PLCs.

Different from the contents posted by personal blogs or programmers, reports published by cyber threat analysts sometimes included detailed analysis of unique features and processes of the exploit. One *report M* stated that the malware operated in a process, which "[i]n the non-recoverable fault state, the CPU has entered an error mode where all remote communications have been stopped, process logic stops execution, and the device requires a physical power cycle to regain functionality." Accompanying the PoC exploits, detailed analysis discussed in security reports could provide further information to both professional researchers and hackers.

Educational materials

This theme involved different kinds of open-access resources that were able to directly teach people hacking-related knowledge. Al-

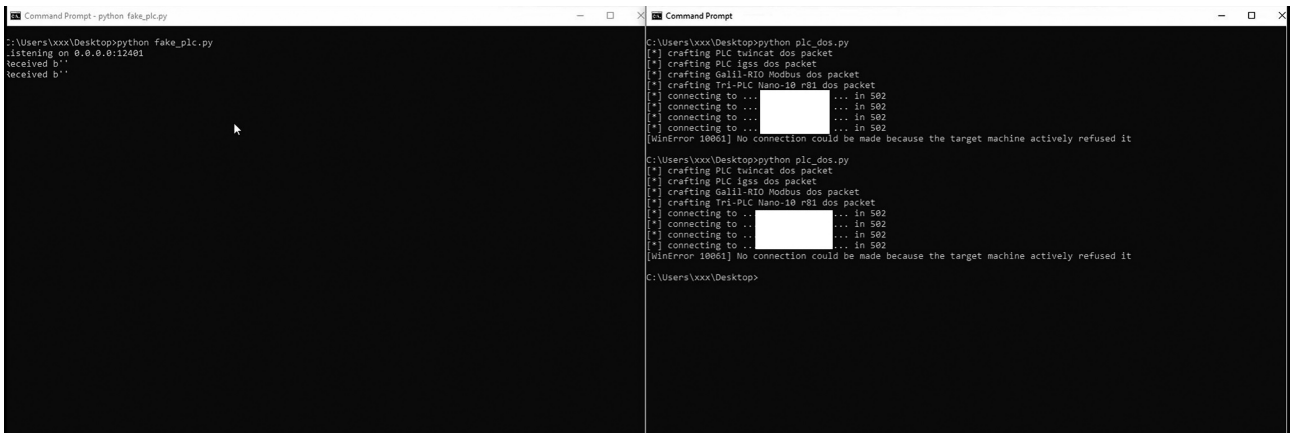


Figure 5: DoS attack toward virtual PLC performed by *User J,* available from *Claes* channel on YouTube

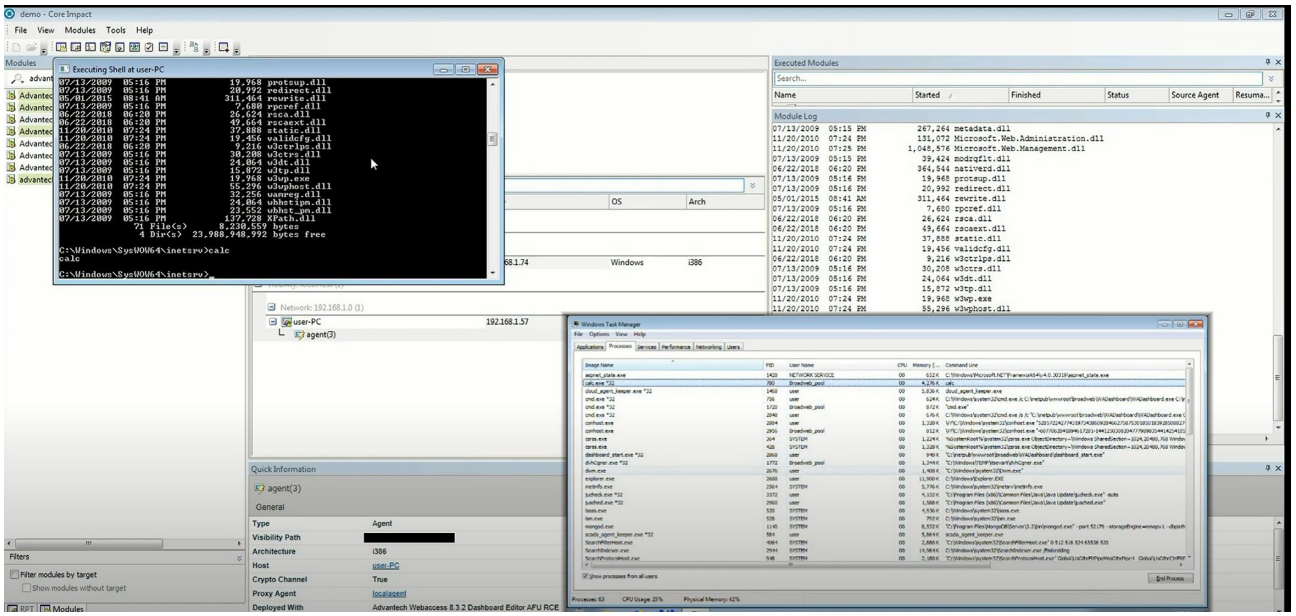


Figure 6: Remote code execution on Advantech WebAccess by ExCraft's *Lab K* (YouTube)

though the frequency of materials related to the theme (13.6%) was relatively lower than the other two main themes identified in the current study, these educational resources provided attackers important insight on how to perform cyber attacks against CIs. A total of two types of materials were identified from the current sample: (1) "how to" tutorials, and (2) training courses.

"How to" tutorials

The overwhelming majority of the data in this category were found in personal blogs or YouTube videos teaching viewers the steps needed to hack into a SCADA system and to develop exploits. This result suggested that attackers were able to find materials providing detailed information with regards to cyberattacks against CIs in the open domain.

A small number of blogs detailed the steps a person needs to recreate a malware or improvise an existing exploit for a cyberattack. For instance, *Researcher N* posted a blog discussing what hackers should do to create an exploit similar to Stuxnet and deploy it to Schneider Modicon PLCs:

> *… Our breakpoint on MyAsmArmStream is reached, and if we follow the arguments in the stack, we can see that the first argument contains a pointer to our ASM source code. Now we will execute the function MyAsmArmStream and see what happens. … So, we disassemble the byte code at the offset of the label DebugLabel2 (offset bytecode + 0 × 90) to ensure that we retrieve the ASM source code. We have recovered our original ASM source code, so MyAsmArmStream is in charge of the compiled processing. Therefore if we hook into this function, we will be able to inject our own "malicious" code.* (*Researcher N*, 2020)

While tutorial blogs tend to include both verbal and imagery explanations to help readers understand the hacking processes, some of the steps were still omitted and not described in these instructional blogs. Slightly different from these tutorials, *User O* decided to teach others how to write buffer overflow exploits with shellcode by uploading videos on YouTube:

```
# msfvenom -p generic/tight_loop --platform windows_86 -f perl -e x86/shikata_ga_nai
# print /x &loop
# $1 = 0x555555558030

open(code, ">exploit.msw");
binmode(code);
$loop =
"\xbb\x3c\x56\x3b\x1e\xd9\xc4\xd9\x74\x24\xf4\x58\x2b\xc9" .
"\xb1\x01\x31\x58\x14\x83\xc0\x04\x03\x58\x10\xde\xa3\xd0" .
"\xe0";

print code $loop;
close(code);
```

Figure 7: Partial PoC code of Modbus buffer overflow (*User L*, EDB)

```
    res = getPLCInfo(s)

    # first write system bits and blocks
    mbtcp_fnc = "\x5a"
    session   = "\x00"
    umas_fnc  = "\x23"
    crc = struct.unpack("<I", res[14:18])[0]
    shifted_crc = crc << 1
    crc = struct.pack("<I", shifted_crc)
    data = "0101100080000000c080f3a0a70000200000".decode('he
x')
    umas = "%s%s%s%s%s" % (mbtcp_fnc, session, umas_fnc, crc,
data)
    send_message(s, umas=umas)

    # get plc info
    getPLCInfo(s)

    # second write system bits and blocks
```

Figure 8: PoC exploit against Schneider Modicon PLC (*Report M,* Talos Intelligence)

> *When you use CAT without parameters, it simply redirects its standard input to the standard output. See like here, you type something in, and it gets reflected out. Now you can chain programs together on one line, for example, with semicolon, so we can first print the output of the exploit, and afterwards CAT is executed, so we can enter new input, and if we group that now with some brackets, and redirect their combined output into the stack level, the exploit will first run and execute a shell, and then CAT will take over and we can simply relay input via the CAT to the shell. … It works! We have an ugly shell, and we can verify our identity with "whoami", or id. So now we escalated privileges to root.* (*User O*, 2016)

In their tutorial, *User O* discussed how individuals could use non-functioning source codes and modify them to a root shell and obtain root access to a target system. In addition to the instructions provided by *User O*, extra source codes and links to programs were also shared in the description section, allowing viewers to obtain copies of the code and practice writing exploits on their own.

Training courses

Advertisements for sessions teaching individuals coding and ethical hacking were offered on open-source websites. Typically, people would need to provide some personal information in order to register for the courses; the majority of the courses only require individuals to provide their full name and valid email address to subscribe to the course packages. Some of the ethical hacking training courses required more information, including the applicant's company names, valid work or university emails, and home addresses to ensure the training was done for legitimate purposes.

For both online training and in-person training courses, many of them are offered only for a limited period of time. These courses would take an individual approximately a week to complete, depending on the schedule and the objectives of the course. For example, one course offered by *security researcher P* was designed to be a 4-day training course held during a virtual Black Hat event. Within this 4-day training schedule, *P* stated that the course was planned to:

> *…teach hands-on penetration testing techniques … [that] will apply directly to systems such as the Smart Grid, PLCs, RTUs, smart meters, building manufacturing, …, SCADA, …, and even IoT. … The course exercises will be performed on a mixture of real world and simulated devices to give students the most realistic experience as possible in a portable classroom setting.* (Black Hat, 2020)

In some cases, some of the courses not only taught individuals how to conduct penetration testing against industrial systems, but

also offered an ethical hacker certification exam upon completion of the course package. *Website Q* stated in its course webpage that,

> *… once you've completed … [the course] and practiced your skills in the labs, you're ready to take the certification exam. … the OSCP exam has a 24-hour time limit and consists of a hands-on penetration test in our isolated VPN network. … A passing exam grade will declare you an Offensive Security Certified Professional (OSCP) … [that] is well-known, respected, and required for many top cybersecurity positions.* (*Website Q*, 2020)

These courses indicated that the training sessions were mainly designed for technology professionals or individuals planning to obtain an ethical hacker certification in order to apply for jobs in the cybersecurity field. Upon learning the skills and techniques required for performing penetration testing on different systems, individuals may attempt to illegally apply these skills and exploit devices used in CI.

Discussion

Open-source data could become a threat toward CI as various OSINT tools could inspire attackers on planning and conducting cyberattacks against targeted providers. The themes uncovered in the current study provided a general overview of the types of useful information discoverable by malicious attackers in open-source surface web domains. Further, the findings in this study supported results from previous OSINT literature, demonstrating that the security risks associated with open-source data from surface web on CI should be addressed [2, 33, 39, 40].

Research questions

Our first research question aimed to examine the types of CI-related data available from various surface web platforms. Through collecting and analyzing data from Google, YouTube, Reddit, and Shodan, three major categories of data were found: (1) CI-related reconnaissance data and information-gathering tools; (2) malware PoC codes; and (3) educational materials relevant to hacker skills training. Reconnaissance data related to CI security and threats, hacking and information-gathering tools, as well as demonstration videos of hacking process against CI devices are commonly shared in various websites and platforms. Malicious individuals looking for malware or vulnerabilities' PoC codes can retrieve data from websites such as EDB, Github, or official government or cybersecurity reports. For offenders, training tutorials and educational courses on hacking and exploit development are retrievable for free or upon registration through personal blogs, video channels, and training sites.

To further understand the value of the data, our second proposed research question explored the potential use of these data in the hands of malicious cyber-attackers. The findings suggest the plausibility of ill-intended hackers using the identified types of data for the purposes of reconnaissance information-gathering, re-creation and improvisation of malware against CI devices, or to learn hacking skills. Using data collected through open-source platforms, offenders can thoroughly research their target, learn adequate hacking skills, and ultimately plan and select the desirable strategy to efficiently achieve their objectives. The availability of PoC codes, on the other side, provide hackers resources to reverse engineer pre-existing malware against industrial devices, improve the programs, and potentially deploy them against CI facilities. Our findings illustrate how reconnaissance and weaponization stages (stages 1 and 2) within the cyber attack kill-chain are the predominating phases where malicious attackers will gather and make use of these CI-related OSINT data [1,2,15,16].

CI-related data, researching, and planning

Information gathered by hackers during the reconnaissance stage of a cyber attack were found to be the most prominent type of data among all the websites analyzed in the sample. The importance of this information has been especially clear in studies related to cybersecurity. The success rate of cyber attacks on ICS systems relies on whether malicious individuals gained sufficient knowledge on the functioning and interactions between cyber, control, and physical layers of systems [15]. Particularly, information gathering during the research phase of the cyberattack was vital and necessary for both planning and later stages of cyber attacks [3].

As observed from the current research, reconnaissance data related to CI facilities on the surface web can be retrieved by applying keyword searches. Close examination of the open-source data revealed that the majority of the materials could provide indirect forms of information with regards to exploitable devices, vulnerabilities, as well as programming algorithms useful in future attacks. In addition, the relatively lower level of skills required to obtain such OSINT data suggest the possibility for novice hackers to take part in complex cyberattacks against CI through the means of data-gathering. Script kiddies, if not the only hackers involved in planned cyberattacks, could become a threat to CI facilities by recklessly researching and providing useful reconnaissance and OSINT data necessary for a successful cyberattack [41].

Our findings corresponded with previous research and demonstrated the negative effect of open-source information exposure relevant to industrial infrastructures [4, 40]. If malicious attackers are indeed searching for vulnerabilities of industrial devices and methods to break into CI facilities, information related to attack strategies such as buffer overflow, man-in-the-middle attacks, denial of service (DoS) attacks, as well as default credentials were often commonly accessible and exploitable [4, 33, 39]. For example, Albataineh and Alsmadi's research recently showed that 18 539 out of the 80 611 active devices returned from Shodan queries were found to use default credentials [33]. Kaspersky ICS CERT also published a report identifying the most popular types of exploits used by hackers, where data relevant to the malware and attack strategies can be easily gathered through OSINT [39]. The processes and effort needed to look for relevant data, as shown in both our findings and previous literature, are not difficult for offenders. In the findings section, all the information related to these common exploits, such as default usernames and passwords, were shown to be searchable and collectable from surface web platforms to help hackers identify vulnerabilities in CI facilities and plan out effective cyberattack strategies. These results can be contextualized in the broader research relevant to data gathering and CI security: researching and using CI-related data where relevant information can be easily obtained through OSINT may make the planning of attacks easier.

Data from websites are not the only sources of information related to cyber-vulnerability and the identification of such exploits. As has been presented in our study, different types of OSINT tools such as Shodan and information-gathering software continued to prove their abilities in providing information to hackers regarding CI facilities. Industrial devices searchable by network scanners are at higher risk of being hacked or exploited by malicious individuals, since sniffing tools such as Shodan, Nmap, and Nessus are capable of gather-

ing details about industrial devices and network communications [1]. For example, Samtani *et al*.'s [4] assessment on SCADA devices using both Shodan and Nessus revealed >500 000 devices discovered by the scanners, providing information on thousands of vulnerable devices and models exploitable by malicious hackers. In other words, rich information returned by network scanners and other tools allow offenders to identify and select the ideal devices and "targets" for the cyber attacks. Playing an important role in the reconnaissance stage within a cyber kill chain, successful vulnerability identification, and planning of a cyberattack using data gathering tools and techniques can essentially lead to a variety of available cybercrime methods suitable for targeted devices.

Although it was a small part of our findings, the discussion of social engineering strategies in conferences and presentations was also a significant factor increasing the risk of cyberattack. With the ever-increasing popularity of social media and online socialization, the gathering and exploitation of open-source personal data is attracting both researchers and offenders' attention. For instance, Hayes and Cappa's [18] study reported series of information on social networking websites useful for hackers during the process of finding desirable social engineering targets. Their finding suggested that exposed individual characteristics such as marital status, level of education, age, as well as political preferences would affect the risk of social phishing [18]. As employees used the internet with greater frequency and posted more personal information on social media websites, more opportunities were provided to hackers in conducting cyberattacks through methods such as social engineering. More relevant to our findings, the discussions of social engineering methods and specific personal profiles of vulnerable targets might aid hackers in the progress of learning social engineering, adopting the strategies, and identifying favorable targets.

PoC codes, re-creation, and improvisation of malware

While we have seen an extensive number of collectable information relevant to data gathering and reconnaissance stages of cyber attacks, a small portion of the OSINT data retrieved for this study also contained PoC codes of malware against CI facilities. The publication of PoC codes found in the study would negatively impact the security of industrial systems; hackers could potentially conduct cyber attacks through the re-creation of the zero-day exploit against unpatched industrial devices.

The starting point of disclosing PoC codes of malware was to consider the positive collaboration and exchange of opinions between information technology (IT) professionals and experts [37–39]. The idea was attractive because as proposed by researchers, such publication would benefit the community and ultimately push for better security strategies and methods to protect CI. In practice, however, researchers suggested that the sharing of these codes in public may provide attackers opportunities to reverse-engineer the source code and use the malware against unpatched devices [35]. Wang *et al*. [40] discovered that fully disclosed PoC exploits online could be turned into active exploitable codes through successful alteration to the exploitable states of the disclosed codes. Since hackers could identify vulnerabilities in CI devices and determine ideal attack strategies using informational OSINT data, the existence of PoC codes in the public could put CI facilities at higher risk by allowing hackers to reverse-engineer and modify the existing malware exploits.

Another underlying issue relevant to the publication of PoC codes is the owners' lack of awareness on cybersecurity and their reluctance to maintain and protect their devices. An example of this is given in Positive Technologies' report of technical analysis and PoC exploits [35]. Despite observing a rising trend in the variety and frequency of cyber attacks against companies and devices, many vendors were reluctant to upgrade the system and patch the vulnerabilities upon the release of the fixes [35]. A total of 1 year after the publication of Positive Technologies' report, statistical data from Kaspersky ICE CERT also found evidence that 32% of all industrial devices in the year of 2019 were exposed to cyber attacks due to unpatched vulnerabilities and outdated programs [39]. In short, industrial devices not receiving required maintenance and fixes are also at higher risk of being hacked.

Educational materials, learning, and hacking skills training

Finally, the exploration of the types of OSINT data available in publicly available surface web platforms has also shown several types of educational materials relevant to hacking, including blogs, tutorials, and courses. A portion of the results (13.6%) in the current study were able to provide tutorials and hacking courses, allowing hackers to learn specific data-gathering strategies and attack methods against targeted devices. One specific type of data we found in the current study, demonstration videos of successful hacks against ICS systems, were yet being discussed in previous research. Although we were unable to locate related literature discussing the effects and usefulness of demonstration videos, we can hypothesize based from criminological learning theories, that hackers may gather and learn programming-related data through watching these demo videos and educational materials. In other words, all educational materials available in public domains may be gathered and are helpful for hackers throughout their learning processes.

An individual's involvement in cybercrimes could be explained by Aker's social learning theory; hackers are more likely to engage in computer crimes when differentially associated with other hackers and imitating their behavior [42]. In addition to the social learning components, more learning opportunities and online resources are presented to hackers, allowing them to obtain the required hacking and data-gathering skills [43, 44]. Further, the availability of such resources drastically decreased the requirement of physical–social interactions between deviant individuals for skill-learning purposes [12].

In line with ideas proposed in previous literature, our findings showed that there is an abundance of hacking-related information circulating in OSINT platforms [44]. As resources and learning opportunities related to hacking increase online, people's engagement in hacking may also increase [42–44]. Increasing accessibility of information on the internet would have a profound impact on people's involvement in computer-based crimes by means of self-empowerment and self-facilitated learning. The findings we presented in the paper could suggest that, other than the more technically advanced hackers, novice hackers can now obtain hacking skills through self-learning and become a threat to CI facilities. While it can be straightforward that online tutorials and educational courses can teach malicious individuals hacking and malware development skills against CI devices, these demo videos may also aid hackers during the self-learning process. It is also possible that hackers could obtain excerpts of the codes presented in the videos and re-create the malware source code through reverse engineering techniques.

Limitations and Future Research

A few potential limitations should be addressed for this study. First, the current sample cannot be generalized to all publicly available information in all social media platforms and search engines. Fur-

ther, our searches may not have returned all available information. Search engines such as Google tend to omit repetitive search results, thus containing only a limited number of websites for each of the keywords. In addition, policies prohibiting illegal information such as content promoting unlawful activities, sharing illegal malware, or posting information related to hacking are often actively implemented among many open access websites [27–29]. A detection of illegal content violating open web policies will result in a removal of said content [29]. For platforms like YouTube or Reddit where the total quantity of results was not provided, we were able to analyze a random selection of the results and saturation was reached. Thus, it seemed reasonable to assume generalizability toward the keyword search results emerged from these open-source websites, not including results from additional keywords and other search engines.

Similar to other research on OSINT data, the authenticity of the textual contents analyzed in the current study may be an issue. Open-source information can be difficult to validate since individuals can easily share any experiences or information on the surface web. In contrast to studies reliant on the analysis of real-life social media and forum data, the sample included in the current study are mainly official announcements, conferences, and open-access white papers published by IT experts. Only a small portion of subjective publications are included in the sample. Hence, the accuracy and validity of the content included in the sample could be considered genuine.

Additionally, it is also important to address the limitation of the data collection method used in the current research. Like all of the qualitative research involving analysis over both textual and video/audio contents online, the manual extraction and analysis of data would increase the time required for data collection, making researcher fatigue a rising issue. Enhancement of data extraction and analysis is particularly needed for OSINT data, as open-source platforms can contain rich information relevant to the topic of interest. To address this issue, it is recommended that researchers use computerized techniques and automated programs in future studies to automatically extract and filter data relevant to the research topic.

Notwithstanding the above limitations, this study has contributed new information that had not been addressed in previous literature. Particularly, we were unable to find any studies discussing the use of videos demonstrating successful attacks toward exploitable vulnerabilities. We hypothesize from the results that the discovery of these videos may serve as educational channels and facilitate ill-intended individuals to learn hacking skills online. The increasing variety of learning materials and individual engagement in the learning progresses may be useful in predicting one's involvement in cyberattack behavior. Our observation indicated the plausibility of applying criminological learning theories to help researchers understand the types of OSINT data useful for hackers to self-learn and to carry out deviant hacking behaviors. Therefore, we propose that future research could investigate the potential impact of demo videos in the field of cybersecurity. Additional studies should also focus on examining the impact and predicting power of social learning in the context of open-source data. Comparing the impact of different types of OSINT resources on individuals might allow researchers to identify materials more likely to affect individuals' cybercrime engagement. Further research on the impact of publicly available PoC codes is also recommended as the potential danger of fully disclosed PoC exploits against CI remains under-studied.

Building off from the current findings, mitigation strategies such as implementation of mandatory cybersecurity training, routine vulnerability assessments toward devices and facilities, as well as enforcement of stronger security measures are recommended to ICS vendors. Moreover, current policies regarding sharable resources in the public domain should also be reviewed with the intention to limit the amount of open access information exploitable by hackers, while balancing people's right to learn and read relevant cybersecurity materials. We should also consider extending our research to other OSINT resources using additional keyword queries and toward additional OSINT platforms. As this current study only acknowledged a partial list of keywords related to CI, more in-depth analysis could also be designed to discover further open-source information useful to attackers in various cyber-attack phases. Furthermore, future analysis on the impact of various OSINT resources on the security of CI sectors is also recommended, as such information will allow cybersecurity technicians to gain better understanding of the vulnerabilities and enforce more effective mitigation strategies to protect CI industry.

Conclusion

The development of technology has allowed society to access almost everything with smart devices and the internet. When the majority of industries and infrastructures are becoming more reliant on electronic systems and the convenience brought by the internet, it also increased cybersecurity risks against these industrial systems [5].

Aiming to understand the types and content of information obtainable to malicious hackers from open-source surface web websites against the CI industry, we analyzed publicly available OSINT resources through a qualitative research method approach. This study mainly focused on CI-related data that was able to either directly or indirectly help hackers throughout different cyberattack kill-chain phases. Within the results, we noticed the increasing amount of information that could be used by malicious attackers against various industrial devices. Particularly, open access cybersecurity information originally published for security professionals with benign intents and educational purposes are now at risk of being used maliciously by hackers. Incorporating data gathered from open-source websites and skills learned from hacking tutorials and courses, malicious hackers could plan out efficient cyberattacks against ideal targets, resulting in disruptions to CI.

The findings from the study also suggested that novice hackers such as script kiddies may pose an increasing threat toward CI facilities in their abilities to facilitate the low-skill data gathering reconnaissance process, as well as the potential to self-learn to become skilled hackers. Although many known cyberattacks against CI facilities are identified to be sponsored by nation states, it is still possible for independent, motivated hackers to use OSINT data and gather information related to CI facilities and devices. Recruiting and hiring of these motivated hackers to continue the data-gathering processes, or purchasing of these OSINT data could be done by governmental agencies with sufficient resources and funding. Further, it is also possible for independent hackers motivated by monetary gain or reputational needs to conduct cyberattacks against CI facilities in the future. Despite the lack of evidence of cyberattacks being launched depending solely on the use of OSINT data, it is still remarkably important for researchers to recognize the possibility of OSINT-based attacks in the future and to identify potential threats at early stages as institutions are becoming more reliant on remote controlling and accessing of CI facilities.

Acknowledgments

This work was supported by the Natural Resources Canada. The funding source was not involved in the study design; collection, analysis, or interpreta-

tion of data; in writing the report; or in the decision to submit the article for publication.

Conflict of Interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Coffey K, Smith R, Maglaras L. *et al.* Vulnerability analysis of network scanning on SCADA systems. *Secur Commun Netw* 2018;**2018**:1–21.
2. Ghafir I, Saleem J, Hammoudeh M. *et al.* Security threats to critical infrastructure: the human factor. *J Supercomp* 2018;**74**:4986–5002.
3. Rodofile NR, Radke K, Foo E. Extending the cyber-attack landscape for SCADA-based critical infrastructure. *Int J Crit Infrastruct Prot* 2019;**25**:14–35.
4. Samtani S, Yu S, Zhu H. *et al.* Identifying SCADA systems and their vulnerabilities on the Internet of Things: a text-mining approach. *IEEE Intell Syst* 2018;**33**:63–73.
5. Quigley K, Roy J. Cyber-security and risk management in an interoperable world: an examination of governmental action in North America. *Soc Sci Comp Rev* 2012;**30**:83–94.
6. Public Safety Canada. *National Strategy for Critical Infrastructure*. Ottawa: Public Safety Canada, 2009. https://central.bac-lac.gc.ca/.item?id=PS4-65-2009-eng&op=pdf&app=Library (10 August, 2020, date last accessed).
7. Chen TM. *Cyberterrorism after Stuxnet*. Carlisle, PA: Strategic Studies Institute, US Army War College, 2014. http://www.jstor.org/stable/resrep11324 (10 August, 2020, date last accessed).
8. Miller B, Rowe DC. A survey SCADA of and critical infrastructure incidents. In: *Proceedings of the First Annual Conference on Research in Information Technology*, New York, NY, 2012. (pp.51–6).
9. National Institute of Standards and Technology. *Supplemental information for the interagency report on strategic U.S. Government engagement in international standardization to achieve U.S. objectives for cybersecurity*. 2015. https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8074v2.pdf . (12 August, 2020, date last accessed).
10. Tariq N, Asim M, Khan FA. Securing SCADA-based critical infrastructures: challenges and open issues. *Proc Comp Sci* 2019;**155**:612–7.
11. Mittal S, Das PK, Mulwad V. *et al.* CyberTwitter: using Twitter to generate alerts for cybersecurity threats and vulnerabilities. In: *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Davis, CA, 2016. (pp.860–7).
12. Kranenbarg MW, Ruiter S, Van Gelder J. Do cyber-birds flock together? Comparing deviance among social network members of cyber-dependent offenders and traditional offenders. *Eur J Criminol* 2021;**18**:386–406.
13. Pastor-Galindo J, Nespoli P, Marmol FG. *et al.* The not yet exploited goldmine of OSINT: opportunities, open challenges and future trends. *IEEE Access* 2020;**8**:10282–304.
14. Nicholson A, Webber S, Dyer S. *et al.* SCADA security in the light of cyber-warfare. *Comp Secur* 2012;**31**:418–36.
15. Hahn A, Thomas RK, Lozano I. *et al.* A multi-layered and kill-chain based security analysis framework for cyber-physical systems. *Int J Crit Infrastruct Prot* 2015;**12**:39–50.
16. Hutchins EM, Cloppert MJ, Amin RM. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill-chains. In: *Proceedings of the 6th International Conference on Information Warfare and Security*, Washington, DC, 2010. (pp.113–25).
17. Samtani S, Chinn R, Chen H. *et al.* Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence. *J Manag Inf Syst* 2017;**34**:1023–53.
18. Hayes DR, Cappa F. Open-source intelligence for risk assessment. *Bus Horiz* 2018;**61**:689–97.
19. Bodenheim R, Butts J, Dunlap S. *et al.* Evaluation of the ability of the Shodan search engine to identify internet-facing industrial control services. *Int J Crit Infrastruct Prot* 2014;**7**:114–23.
20. Chen Y, Lian X, Yu D. *et al.* Exploring Shodan from the perspective of industrial control systems. *IEEE Access* 2020;**8**:75359–69.
21. Jagatic TN, Johnson NA, Jakobsson M. *et al.* Social phishing. *Commun ACM* 2007;**50**:94–100.
22. Green B, Prince D, Busby J. *et al.* The impact of social engineering on industrial control system security. In: *Proceedings of the 1st ACM Workshop on Cyber-physical Systems – Security and/or Privacy*, New York, NY, 2015. (pp.23–9).
23. Huber M, Kowalski S, Nohlberg M. *et al.* Towards automating social engineering using social networking sites. In: *Proceedings of the 2009 International Conference on Computational Science and Engineering*, Vancouver, 2009;**3**:117–24.
24. Mansfield-Devine S. Critical infrastructure: understanding the threat. *Comp Fraud Secur* 2018;**7**:16–20.
25. Edwards M, Larson R, Green B. *et al.* Panning for gold: automatically analysing online social engineering attack surfaces. *Comp Secur* 2017;**69**:18–34.
26. Kalpakis G, Tsikrika T, Cunningham N. *et al.* OSINT and the Dark Web. In: Akhgar B., Bayerl P., Sampson F. (eds.), *Open-Source Intelligence Investigation: From Strategy to Implementation*. 2016. (pp. 111–32). Cham: Springer International Publishing.
27. Google Search Help. Policies for content posted by users on Search. Google. 2021. https://support.google.com/websearch/answer/7408270?hl=en (14 December, 2021, date last accessed).
28. YouTube Help. Harmful or dangerous content policy. Google. 2021. https://support.google.com/youtube/answer/2801964?hl=en&ref_topic=9282436 (14 December, 2021, date last accessed)
29. Reddit. Reddit content policy. Reddit. 2021. https://www.redditinc.com/policies/content-policy (14 December, 2021, date last accessed).
30. Tor. Censorship. Tor Project. 2021. https://support.torproject.org/censorship/ (16 December, 2021, date last accessed).
31. DuckDuckGo. Privacy. *DuckDuckGo*. 2021. https://duckduckgo.com/privacy (16 December, 2021, date last accessed).
32. Palys T, Atchison AJ. Text, image, audio, and video: making sense of non-numeric data. In: *Research Decisions: Quantitative, Qualitative, and Mixed Method Approaches*. 5th edn. 2013. (303–32). Toronto: Nelson Education.
33. Albataineh A, Alsmadi I. IoT and the risk of internet exposure: Risk assessment using Shodan queries. In: *Proceedings of the 2019 IEEE 20th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, Washington, DC, 2019. (pp.1–5).
34. Cartagena A, Rimmer G, Van Dalsen T. *et al.* Privacy violating open-source intelligence threat evaluation framework: a security assessment framework for critical infrastructure owners. In: *Proceedings of the 2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, 2020. (pp.0494–9).
35. Positive Technologies. *Cybersecurity Threatscape: Q4 2018*. 2018. https://www.ptsecurity.com/upload/corporate/ww-en/analytics/Cybersecurity-threatscape-2018-Q4-eng.pdf (18 August, 2020, date last accessed).
36. Rehg J, Sartori G. Instructional algorithms enhance student understanding of PLC ladder logic programming. In: *Proceedings of the 2010 Annual Conference and Exposition*, Louisville, KY, 2010. (pp. 15.751.1–15.751.13).
37. Peterson D. Project Basecamp at S4. *Dale Peterson*. 2012. https://dale-peterson.com/2012/01/19/project-basecamp-at-s4/ (16 August, 2020, date last accessed).
38. S4 Events. Project Basecamp – PLC Hacking Intro. *YouTube*. 2016. https://www.youtube.com/watch?v=BKJje3Ram2I&t=12s (16 August, 2020, date last accessed).
39. Kaspersky ICS CERT. Threat landscape for industrial automation systems: H2 2019. 2020. https://ics-cert.kaspersky.com/media/KASPERSKY

_H22019_ICS_REPORT_FINAL_EN.pdf (18 August, 2020, date last accessed).
40. Wang Y, Wu W, Zhang C. *et al*. From proof-of-concept to exploitable. *Cybersecur* 2019;**2**:1–25.
41. Verton D. Black hat highlights real danger of script kiddies. Computerworld. 2001. https://www.computerworld.com/article/2581986/black-hat-highlights-real-danger-of-script-kiddies.html (24 December, 2021, date last accessed).
42. Holt TJ, Burruss GW, Bossler AM. Social learning and cyber-deviance: examining the importance of a full social learning model in the virtual world. *J Crime Just* 2010;**33**:31–61.
43. Dearden TE, Parti K. Cybercrime, differential association, and self-control: knowledge transmission through online social learning. *Am J Crim Just* 2021;**46**:1–21.
44. Goldsmith A, Brewer R. Digital drift and the criminal interaction order. *Theor Criminol* 2015;**19**:112–30.