Cyber Reconnaissance Techniques

Article *in* Communications of the ACM · February 2021

DOI: 10.1145/3418293

Wojciech Mazurczyk
Warsaw University of Technology
255 PUBLICATIONS **4,865** CITATIONS

SEE PROFILE

Luca Caviglione
Italian National Research Council
194 PUBLICATIONS **2,484** CITATIONS

SEE PROFILE

DOI:10.1145/3418293

The evolution of and countermeasures for ...

BY WOJCIECH MAZURCZYK AND LUCA CAVIGLIONE

Cyber Reconnaissance Techniques

ALMOST EVERY DAY, security firms and mass media report news about successful cyber attacks, which are growing in terms of complexity and volume. According to Industry Week, in 2018 spear-phishing and spoofing attempts of business emails increased of 70% and 250%, respectively, and ransomware campaigns targeting enterprises had an impressive 350% growth.[19] In general, economic damages are relevant, as there is the need of detecting and investigating the attack as well as restoring the compromised hardware and software.[15] To give an idea of the impact of the problem, the average cost of a data breach has risen from $4.9 million in 2017 to $7.5 million in 2018.[19] To make things worse, attackers can now use a wide range of tools for compromising hosts, network appliances and Internet of Things (IoT) devices in a simple and effective manner, for example, via a Crime-as-a-Service business model.[11]

Usually, each cyber threat has its own degree of sophistication and not every attack has the same goal, impact, or extension. However, the literature agrees that an attack can be decomposed into some general phases as depicted in Figure 1. As shown, the Tao of Network Security Monitoring subdivides the attacks in to five stages[6] and the Cyber Kill Chain in to seven stages,[26] whereas the ATT&CK framework proposes a more fine-grained partitioning.[27] Despite the reference model, the first step always requires gathering information on the target and it is commonly defined as "*reconnaissance*." Its ultimate goals are the identification of weak points of the targeted system and the setup of an effective attack plan.

In general, reconnaissance relies upon a composite set of techniques and processes and has not to be considered limited to information characterizing the target at a technological level, such as, the used hardware or the version of software components. Attackers also aim at collecting details related to the physical location of the victim, phone numbers, names of the people working in the targeted organizations and their email addresses. In fact, any bit of knowledge may be used to develop a software exploit or to reveal weaknesses in the defensive systems.

Unfortunately, the evolution of the Internet, the diffusion of online social networks, as well as the rise of services for scanning smart appliances and IoT

» key insights

- **An attack can be decomposed into some general phases. The first step always requires gathering information on the target, a.k.a. "reconnaissance."**
- **There is a plethora of reconnaissance techniques available for an attacker and many of them do not even require a direct contact with the targeted victim.**
- **Counteracting reconnaissance attempts must be viewed within the framework of the "arms race" between attackers and defenders.**
- **Defenders appear to be a step back with respect to attackers. Countermeasures should aim to: strengthen training, enforce proactive approaches, explore cyber deception as a defense tool, engineer reconnaissance-proof-by design services, and rethink the privacy concept.**

IMAGE BY DESTROLOVE

Figure 1. The most popular reference models used to decompose a cyber attack into phases.

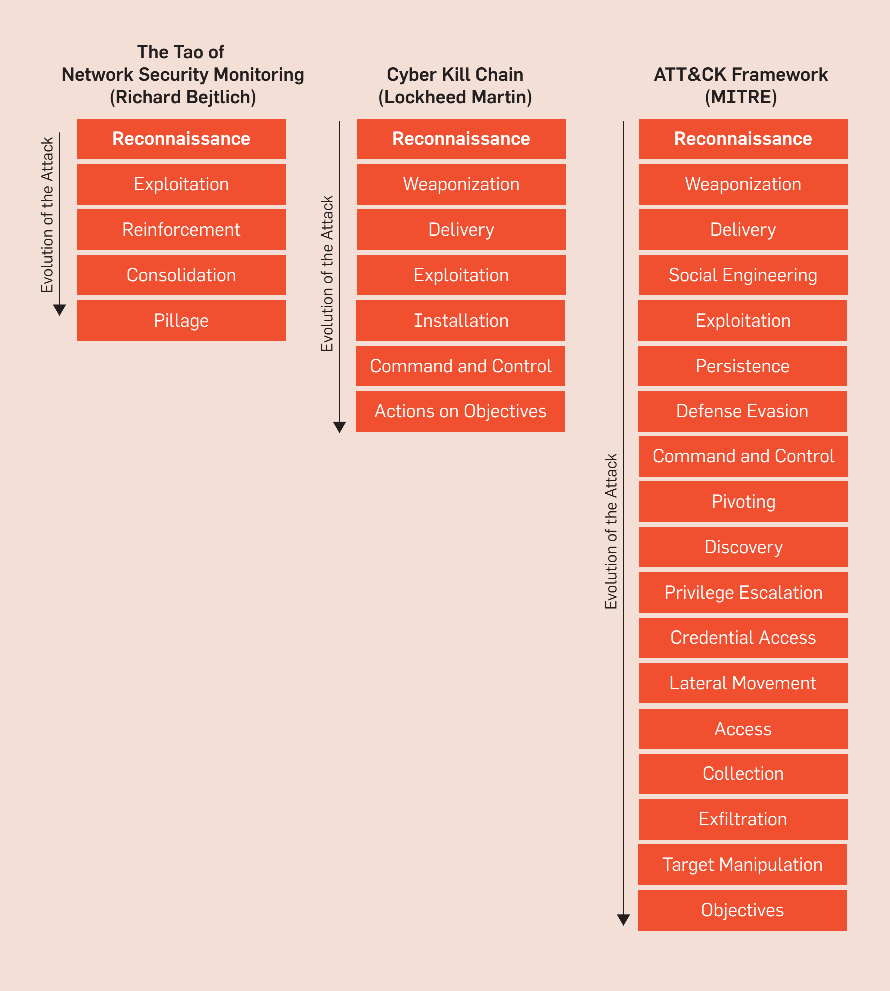
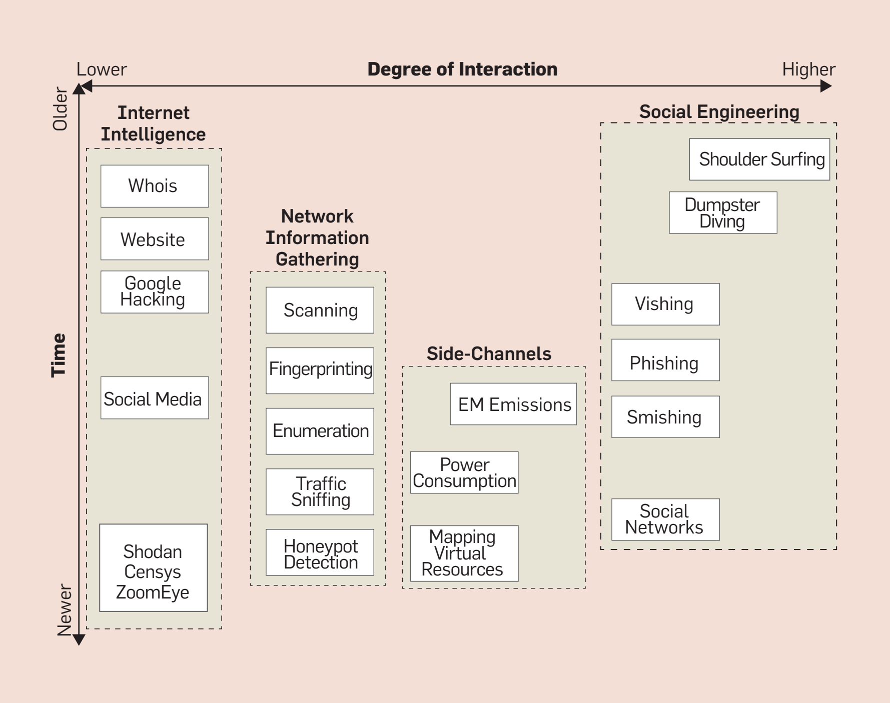


Figure 2. Classification of the reconnaissance techniques and their organization according to the time of appearance and the required degree of interaction with the victim.



nodes, lead to an explosion of sources that can make the reconnaissance phase quicker, easier, and more effective. This could also prevent contact with the victim or limit its duration, thus making it more difficult to detect early and block reconnaissance attempts. Therefore, investigating the evolution of techniques used for cyber reconnaissance is of paramount importance to deploy or engineer effective countermeasures. Even if the literature provides some surveys on some specific aspects of reconnaissance (see, for example, network scanning[8] and techniques exploiting social engineering[31]) the knowledge is highly fragmented and a comprehensive review is missing. In this perspective, this paper provides a "horizontal" review of the existing reconnaissance techniques and countermeasures, while highlighting emerging trends.

In this article, we introduce the classification and the evolution of the most popular reconnaissance methods. Then, we discuss possible countermeasures and present some future directions.

Classification and Evolution

In order to illustrate the most important cyber reconnaissance techniques and portrait their evolution, we introduce the following taxonomy composed of four classes:

- *Social Engineering:* It groups methods for collecting information to deceive a person or convincing him/her to behave in a desired manner.
- *Internet Intelligence:* It groups methods taking advantage of information publicly available in the Internet including databases accessible via the Web.
- *Network Information Gathering:* It groups methods for mapping the network (or computing) infrastructure of the victim.
- *Side-Channels:* It groups methods exploiting unintended information leaked by the victim.

Each class accounts for a given "degree of interaction" with the victim, with the wide acceptation of how tight the coupling with the source of information should be for the purpose of the reconnaissance. For instance, reading the computer screen requires to be near the victim, thus potentially having a physical interaction, whereas scanning his/her network can be done remotely. In addition, some side-channels exploit a measurement that entails to be physically in a proximity to the target (for example, to measure the intensity of an electromagnetic field or the temperature of a heat source), while retrieving data from a social network does not require interacting with an asset run or owned by the victim itself.

Clearly, planning sophisticated attack campaigns or bypassing multiple security perimeters (for example, virtualized services deployed within a De-Militarized Zone) could require combining methodologies belonging to different classes. The longer the attacker actively interacts with the target, the higher the chance the attempt could be detected and neutralized. Unfortunately, the advent of social media, the progressive digitalization of many processes and workflows (as it happens in Industry 4.0 or in the smart-* paradigm), as well as the increasing pervasive nature of search engines, make the collection of data quicker and more effective. In this vein, Figure 2 proposes the taxonomy of reconnaissance techniques and it also emphasizes their temporal evolution and the required degree of interaction with the victim. We underline that the figure is intended to locate in time when methods firstly appeared and not how long they have been used (actually, the majority still is in the toolbox of attackers).

We now review the most important reconnaissance techniques proposed in the literature and observed in the wild, which are summarized and further commented in the sidebar "Examples of Reconnaissance Techniques and Sources."

Social engineering is probably the oldest family of techniques used for reconnaissance and it is extraordinarily effective as it exploits the weakest link in security: humans. In essence, social engineering tries to manipulate and deceive victims by misusing their trust and convince them to share confidential information or to perform activities that can be useful to the attacker, for example, download and install a keylogger. It can also significantly decrease the time needed to gather information and often requires minimal or none technical skills.[31] The literature proposes several taxonomies for social engineering attacks,[31] but the simplest subdivision considers two main groups: *human-based* requiring a direct or in person interaction, and *technology-based* where the physical presence of the attacker is not needed. Human-based techniques are the oldest and include methods like impersonation, dumpster diving or shoulder surfing. Even if still used, technology-based mechanisms today appear to be more popular and include methods like phishing and spam, or for tricking the user to install malware by using pop-ups and ad hoc crafted email.

Examples of Reconnaissance Techniques and Sources

Social Engineering

Shoulder surfing: techniques where the attacker tries to determine confidential data by looking over the shoulders of the victim.

Dumpster diving: the practice of obtaining information from discarded material, such as documents, components of computing devices like hard drives and memory cards.

Phishing/Vishing/Smishing: the attacker tries to mislead the victim by impersonating a trustworthy entity by using email, VoIP, and Short Message Service.

Social Networks: the attacker utilizes social networks (for example, Facebook, LinkedIn, and Twitter) for gathering personal data or persuading the victim to reveal sensitive information or accomplish certain actions.

Internet Intelligence

whois/rwhois: databases providing information about IP address range and Autonomous Systems used by the victim.

Website: HTML pages can contain a very large and composite set of data. For the case of corporate websites, available information concerns employees, contact details, position within the organization, just to mention some. Comments left in HTML are another valuable source of information.

Google Hacking (Google Dorking): techniques utilizing advanced operators of Google to reveal potential security vulnerabilities and/or configuration errors of hardware and devices managed by the victim.

Social Media: a source of reconnaissance data where an attacker can collect personal information about the victim in order to learn, for instance, his/her habits, hobbies, likes and dislikes, with the aim of creating a more complete profile of the targeted person.

Shodan/Censys/ZoomEye: specific search engines indexing detailed technical data about different types of devices and network appliances.

Network Information Gathering

(Port) Scanning: methods for probing devices to establish whether on the targeted host there are open ports and exploitable services.

(OS/application) Fingerprinting: techniques for recognizing the operating system and/or applications utilized on the targeted device. A host can be stimulated with certain network traffic and replies are analyzed to guess the OS and/or installed applications.

(Network/Device) Enumeration: the systematic process for discovering hosts/servers/devices within the targeted network that are publicly exposed by the victim.

Traffic Sniffing: an attacker infers information about the victim network by collecting (sniffing) traffic or via monitoring tools.

Honeypot Detection: a set of techniques allowing the attacker to recognize whether the compromised machine is real or virtual. Typically, such methods rely on the detailed analysis of the behavior of the breached host (execution delays) or network configurations (MAC address, ARP and RARP entries, and so on).

Side-Channels

EM Emissions/Power Consumption: side-channels can be used to infer the signals leaked from screens, printers, or keyboards, to retrieve sensitive information. The most relevant physical quantities observed to set the side-channels are electromagnetic emissions or the power consumption of targeted devices.

Mapping Virtual Resources: side-channels are used to map a cloud infrastructure in order to establish if services are virtualized/containerized or to perform other types of attacks like co-resident threats. Typically, this class of side-channels operates in a completely remote manner.

Another important goal of social engineering is to get information about the victim or enrich (uncomplete) bits of information gathered with other techniques, for example, compile a custom dictionary to force the password for a username/email observed on a website. With the high exposure of people to several communication channels and the variety of social media services, an attacker has a wide array of opportunities to craft reconnaissance campaigns, as he/she can use face-to-face, telephone/VoIP, or instant messaging services, as well as online scams and fake identity attacks on online social networks. Such risks are exacerbated by the Bring Your Own Device paradigm, which makes it more difficult for an enterprise to control laptops, phones, and smart devices of their employees or to enforce access rules to shared resources like workspaces, wikis, forums, and websites.

Internet intelligence. Searching for publicly available information in the Internet is probably the first step that any attacker performs. The number of sources that can now be queried makes it possible to retrieve a huge amount of apparently insignificant fragments, which can become very informative if properly combined. In this perspective, Internet intelligence is the "offensive" subgroup of Open Source INTelligence (OSINT) and it is specialized and limited to the information available on the Internet and its services, such as the Web, public databases, specialized scanning services to map IoT nodes, and geographical or geo-referenced sources. Fortunately, the General Data Protection Regulation partially mitigated such a risk since the access to many public databases within the European Union is restricted. Internet Intelligence can be also used to perform passive *footprinting*, that is, the collection of publicly available information to identify a hardware and/or software infrastructure. In the following, we will discuss the main usage trends of Internet Intelligence.

Web sources. The increasing pervasive nature of the Web and the evolution of search engines surely added new and powerful options into the toolbox of attackers. Typically, a reconnaissance campaign starts from the website of the victim. In this way, the attacker can gather important data like employee names, email addresses, telephone numbers or the physical address of the target, which can be used to perform social engineering or drive other threats. Personal bits of information can be also "fused" and enriched with data collected from online social networks. For instance, upon retrieving the hierarchy of the company and the list of the employees, the attacker can move to Facebook or LinkedIn to launch phishing or vishing attacks.[12]

The increasing pervasive nature of the Web and the evolution of search engines surely added new and powerful options into the toolbox of attackers.

Search engines are central for Internet intelligence, since they can limit the interaction between the attacker and the victim, thus making the data gathering phase difficult to detect. For instance, the attacker might craft some scripts to perform screen scraping directly from a website close to the victim hence leaving some traces in the log of the Web server. However, cached versions of the webpage provided by services like Google or the Internet Archive[a] can be used to avoid traces of the reconnaissance attempt.

Apart from details that can be retrieved in an organic manner from indexed pages (for example, hobbies, owned books and records or visited stores), search engines can be also used to perform more fine-grained intelligence activities. Google Hacking[23] is one of the most popular techniques and it exploits advanced operators to perform narrow and precise queries mainly to reveal security breaches or configuration errors. For instance, the attacker can use operators like "*inurl*" to search within URLs. Google can then be queried with "*inurl:/hp/device/this.lcdispatcher*" to discover details on a printer model to reveal potential vulnerabilities or search for a precooked exploit. Another possible reconnaissance mechanism mixes the aforementioned technique for "Googling the Internet," that is, using search engines to gather information on endpoints involved in a communication without the need of collecting or analyzing network traffic.[38]

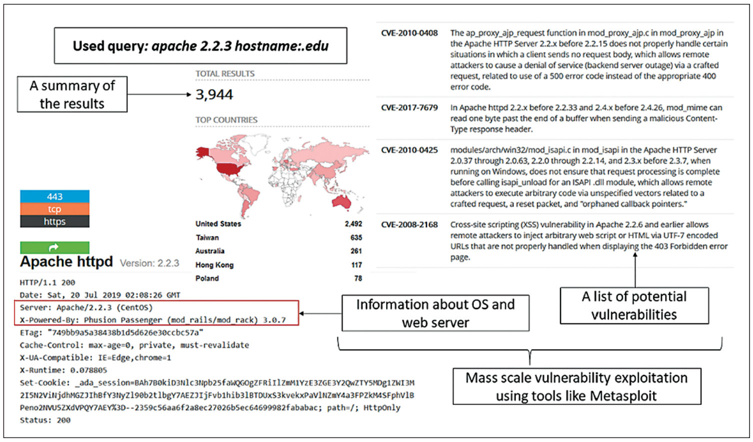
Public databases and sources. The variety of public records available online is another important source of information. In fact, every IP address and

[a] https://archive.org/web/

domain name should be registered in a public database, which can also contain a contact address and a telephone number. Some hints on the "layout" of the network of the victim can be inferred without needing to directly scan hosts or appliances. By querying the American Registry for Internet Numbers,[b] it is possible to obtain the complete block of IP addresses assigned to the target. The Domain Name System can provide a wealth of details on the adopted addressing scheme and naming strategy. Other sources used for reconnaissance are the *whois* and *rwhois*,[c] which can provide IP address blocks and details on the autonomous system of the victim.

Public scanning services. As hinted, a large part of the success of an attack depends on identifying vulnerabilities within the targeted network/system. Until few years ago, this required performing a direct scan toward hosts, network devices, and software components or being able to collect network traffic, for example, via sniffers. To mitigate the direct exposure, a possible technique uses a botnet of zombies, that is, a network of compromised hosts under the control of the attacker. Zombies can then be used as proxies.[13,16] Even if this approach may prevent to trace back the source of the scan/attack, still the attempt can be spotted or hindered. In this vein, a recent trend changed the situation, especially if the reconnaissance campaign targets IoT devices or smart settings like Heating, Ventilation and Air Conditioning. In fact, the availability of tools like Shodan,[d] Censys,[e] and ZoomEye[f] imposed a paradigm shift to reconnaissance. Roughly, such services automatically scan the whole IPv4 public addressing space in a distributed and random manner and offer the obtained knowledge (for example, used hardware, open ports, or types of service delivered) via search-engine-like interfaces or ad-hoc Application Programming Interfaces. See the sidebar "Example of Shodan Query and Related Intelligence" for an example usage of Shodan for Internet intelligence.

Example of Shodan Query and Related Intelligence



Used query: *apache 2.2.3 hostname:.edu*

A summary of the results

TOTAL RESULTS

3,944

TOP COUNTRIES

United States	2,492
Taiwan	635
Australia	261
Hong Kong	117
Poland	78

Apache httpd Version: 2.2.3

```
HTTP/1.1 200
Date: Sat, 20 Jul 2019 02:08:26 GMT
Server: Apache/2.2.3 (CentOS)
X-Powered-By: Phusion Passenger (mod_rails/mod_rack) 3.0.7
ETag: "749bb9a5a38438b1d5d626e30ccbc57a"
Cache-Control: max-age=0, private, must-revalidate
X-UA-Compatible: IE=Edge,chrome=1
X-Runtime: 0.078805
Set-Cookie: _ada_session=BAh7B0kiD3Nlc3Npb25faWQGOgZFRiIlZmM1YzE3ZGE3Y2QwZTY5MDg1ZWI3M
2I5N2ViNjdhMGZJIhBfY3NyZl90b2tlbgY7AEZJIjFvb1hib3lBTDUxS3kvekxPaVlNZmY4a3FPZkM4SFphVlB
Peno2NVU5ZXdVPQY7AEY%3D--2359c56aa6f2a8ec27026b5ec64699982fababac; path=/; HttpOnly
Status: 200
```

CVE-2010-0408	The ap_proxy_ajp_request function in mod_proxy_ajp.c in mod_proxy_ajp in the Apache HTTP Server 2.2.x before 2.2.15 does not properly handle certain situations in which a client sends no request body, which allows remote attackers to cause a denial of service (backend server outage) via a crafted request, related to use of a 500 error code instead of the appropriate 400 error code.
CVE-2017-7679	In Apache httpd 2.2.x before 2.2.33 and 2.4.x before 2.4.26, mod_mime can read one byte past the end of a buffer when sending a malicious Content-Type response header.
CVE-2010-0425	modules/arch/win32/mod_isapi.c in mod_isapi in the Apache HTTP Server 2.0.37 through 2.0.63, 2.2.0 through 2.2.14, and 2.3.x before 2.3.7, when running on Windows, does not ensure that request processing is complete before calling isapi_unload for an ISAPI .dll module, which allows remote attackers to execute arbitrary code via unspecified vectors related to a crafted request, a reset packet, and "orphaned callback pointers."
CVE-2008-2168	Cross-site scripting (XSS) vulnerability in Apache 2.2.6 and earlier allows remote attackers to inject arbitrary web script or HTML via UTF-7 encoded URLs that are not properly handled when displaying the 403 Forbidden error page.

Information about OS and web server

A list of potential vulnerabilities

Mass scale vulnerability exploitation using tools like Metasploit

Similarly to search engines used to index the Web, also in this case, attackers can gather data without directly contacting the targeted device and compile a list of potential targets/victims in a quick and easy manner: literature often defines this as "*contactless active reconnaissance.*"[29]

A recent trend in contactless active reconnaissance combines different publicly available sources. As an example, data collected via Censys can be merged together with the National Vulnerability Database[g] to improve the accuracy of discovering known vulnerabilities.[29]

Network information gathering. When the data publicly available is not sufficient, the attacker needs to directly interact with the infrastructure of the victim. The most popular class of techniques is the one named "*network scanning*" and enables to map a remote network or identify the used operating systems and applications. Typically, network scanning techniques are divided in two main groups: passive and active.[8] In passive scanning, the attacker infers information about the network by monitoring traffic. To this aim, *sniffers* can be deployed to capture and inspect flows, and the most popular tools are *tcpdump*[h] and *Wireshark*.[i] This may also require "mirroring" ports of a network appliance in order to duplicate the traffic. Instead, in active scanning, information is collected by intentionally generating and sending specific packets (also called *probes*) to the network device under investigation and by analyzing its responses. We point out that while performing scanning, attackers should stay "under the radar" to prevent detection due to anomalous traffic. For example, generating too many ICMP packets or incomplete TCP connections can be spotted with Intrusion Detection Systems (IDS) and firewalls.[8] Scanning can be done at different levels of the protocol stack. Here, we present the most popular reconnaissance techniques for network information gathering grouped according to their scope.

Network and device enumeration. Two important parts of activities related to network information gathering, are *network enumeration* for discovering hosts and servers and *device enumeration* for identifying IoT nodes and other devices that are exposed by the victim. Despite using services like Shodan, the attacker may need to "manually" search for devices, for instance, due to the use of private IPv4 addressing schemes or to check the consistency of earlier information. The enumeration of network elements and devices is usually performed via

[b] www.arin.net
[c] whois.icann.org/en
[d] www.shodan.io
[e] censys.io
[f] www.zoomeye.org
[g] nvd.nist.gov
[h] www.tcpdump.org
[i] www.wireshark.org

traffic analysis. However, the increasing diffusion of wireless technologies like WiFi, Bluetooth, and ZigBee to connect smart things like lightbulbs, various sensors, intelligent sockets and locks, makes the advent of a new form of techniques *à la* wardriving (that is, searching and marking for wireless signals for future exploitation). For instance, due to an improper configuration of wireless access points, the electromagnetic signal may be "leaking" outside the physical perimeter controlled by the victim, hence the malicious entity can expand his/her potential attack surface. For the purpose of scanning such a balkanized technological space, tools especially designed for IoT reconnaissance are becoming available. For instance,[33] proposing a passive tool for scanning multiple wireless technologies. An interesting idea is the use of the observed traffic to go beyond the enumeration of devices by classifying the type of the IoT node (for example, a camera or a smart speaker) and its state (for example, a smart switch is turned on or off). This can endow the attacker with very precise reconnaissance information.

Port scanning and fingerprinting. Methods defined as *port scanning* are designed to probe devices to determine whether there are open ports and exploitable services. Even if the literature abounds of methods, the most popular take advantage of the different behaviors of the three-way-handshake procedure of the TCP. Port scanning can then discover whether a remote TCP port is open by trying to send SYN/ACK packets, establishing a complete transport connection, or abort the process in the middle.[8] Its main limitation is the need to maintain a large amount of TCP connections, thus causing transmission bottlenecks or exhausting the resources of the used machine. Consequently, the scanning rate is decreased and the reconnaissance attempt could be detected. A recent trend exploits distributed frameworks able to reduce both the scanning time and anomalous resource usages that could lead to identifying the attacker.[25]

Scanning can be also used to recognize the guest OS or the applications available in the target nodes. This is known as *fingerprinting* and may be implemented both via active and passive methods. For the case of OS fingerprinting, the main technique exploits the fact that the network stack of each OS exhibits minor differences when replying to well-crafted probe packets (for example, the initial sequence number of the TCP segments, the default TTL value for ICMP packets, among others).[8] Such artifacts can be utilized to remotely determine the type and version of the OS of the inspected device. Application fingerprinting uses a slightly different technique. In this case, the attackers take advantage of a "banner," which is a sort of preamble information that a server-side application sends before accepting a client. By stimulating a host with connection requests, they can harvest banners to reveal details on the active applications and services (for example, the version of the software can be used to determine the known vulnerabilities). A typical tool used for active scanning in network information gathering is *nmap*.[j]

Application-level reconnaissance. The class of techniques named *application-level reconnaissance* is recently gaining attention, especially to infer some high-level features of the targeted host. To this aim, the attacker can utilize scanning tools to reveal certain weak points of the victim network. Possible examples of such tools are commercial suites like Nessus,[k] Acunetix,[l] and Vulners[m] or opensource solutions like IVRE[n] and Vega.[o] Another idea exploits probes to quantify the degree of protection of the victim. In this case, the attacker can use the timing of responses obtained to understand whether an antivirus is working on the targeted machine or if its signatures are updated.[2]

Honeypot detection. Honeypots are increasingly used to collect information on malware to organize suitable defense techniques or counterattacks, especially in case of botnets. From an attacker point of view, they represent a hazard since they can be used to disseminate incorrect information, thus (partially) voiding the reconnaissance phase. Therefore, being able to detect confinement in a virtual/fictitious space is a core skill expected for the development of successful threats.[21] To this aim, the attacker could check for the presence of TUN/TAP devices or specific entries in the ARP cache in order to have signatures to discriminate between real or virtual settings.[17] Another mechanism takes advantage of the "fair" behavior of the honeypot, which impedes the node to harm a third-party entity. Thus, the attacker can try to compromise a host and launch some offensive patterns. According to the outcome, he/she can understand whether the node is real or fictitious.[39]

Side-channels. Firstly envisaged by Lampson,[22] the term *side-channel* usually defines attacks to deduce sensitive information by abusing unforeseen information leakages from computing devices. An interesting research direction started in the 1990s[20] with physical side-channels targeting cryptographic algorithms and their implementation. In essence, by inspecting apparently unrelated quantities, for example, the time needed to encrypt a message, the power consumed by a host or the electromagnetic field produced by the CPU of the device, attackers were able to infer information on the used algorithms and keys, thus making it feasible to exfiltrate encryption keys or conduct probabilistic guesses. Thus, in its original vision, a side-channel required a high degree of interaction with the victim.

This class of techniques is *en vogue* again especially for reconnaissance purposes. For instance, it has been proven that information or signals leaked from screens,[14] printers,[4] and keyboards[7] can be used to retrieve login credentials or cryptographic keys. Other types of side-channels are becoming increasingly used, especially those allowing the attacker to control sensors located in close proximity of the target or to infer keyboard inputs on touchscreens,[34] for example, to exploit fingerprints left by user to guess the used unlock patterns or the PIN code.[3] Owing to the high interconnected and virtual nature of modern hardware and software, side-channels attacks

[j] nmap.org
[k] www.tenable.com/products/nessus
[l] www.acunetix.com
[m] vulners.com
[n] ivre.rocks
[o] subgraph.com/vega

can be also operated in a completely remote manner, thus preventing the contact with the victim. For instance, they can be used to map a cloud infrastructure to understand whether services are virtualized or containerized or to perform cache-timing attacks.[30] To sum up, the use of a side-channel is a double-edged sword as it could require some physical proximity and this may increase the risk of exposure of the attacker, thus the value of the obtained information should be carefully evaluated.

Countermeasures

As has already happened in many other fields of cybersecurity, counteracting reconnaissance must be viewed within the framework of the "arms race" between attackers and defenders. Unfortunately, due to the availability of a composite amount of techniques, it is very difficult to completely prevent an attacker from inspecting a target. Over the years, countermeasures evolved and Figure 3 portraits a classification (also in this case, techniques have been located in the graph according to their estimated initial appearance).

As depicted, the evolution in the development of countermeasures experienced three main époques. In the earliest, the prime method aims at *training and raising awareness* of users as to reduce the effectiveness of social engineering or prevent the leakage of sensitive information. To complete this, constant auditing/monitoring campaigns of the information publicly available in the Internet should be performed on a regular basis. The paradigm shift happened when the design of countermeasures moved from considering primarily the technology rather than the human. The first wave deals with *reactive countermeasures* and aims at directly responding to a specific reconnaissance technique, for instance, scanning or sniffing. The more recent trend deals with *proactive countermeasures*: in this case, the attacker is disturbed or hindered on a constant basis, for example, by deliberately disseminating misleading data.

Human-based countermeasures. To mitigate the bulk of information that can be gathered via social engineering, including those available in the Internet, an effective approach aims at reducing the impact of individuals by proper training and education.[35] Specifically, training may limit the exposure to social engineering techniques by explaining to users what kind of information can be publicly shared and how. Training can be also beneficial for technical staff that can learn the tools used by an attacker to reveal security breaches and design workarounds.

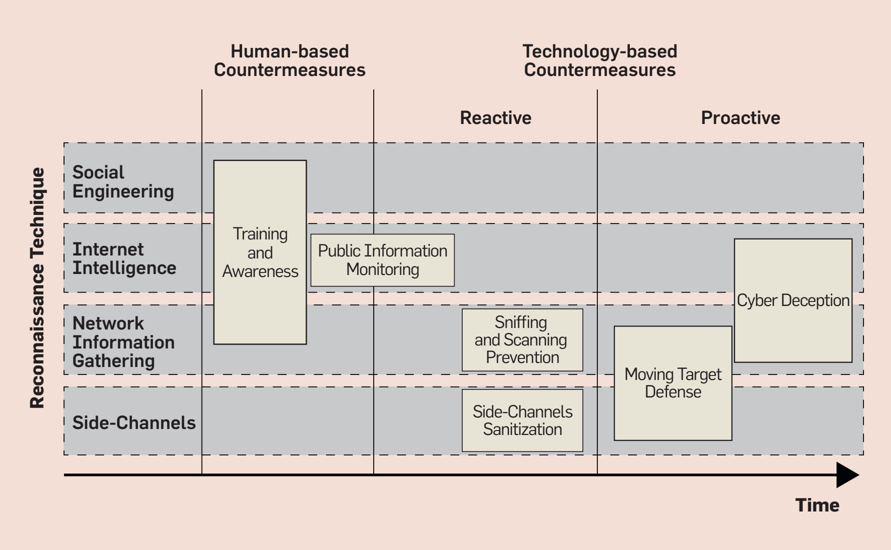
In parallel, security experts should perform public information monitoring on a continuous basis, that is, perform a sort of "protective" OSINT. Obtained data can be used again to instruct users and technicians. More importantly, public information monitoring can help assess the degree of security of the target, sanitize data leaks, as well as feed more sophisticated countermeasures.[18]

Reactive technology-based countermeasures. As hinted, reactive countermeasures are the direct response against reconnaissance attempts, including those exploiting side-channels. The main limitation of the approach is that if the threat evolves in time, the defensive mechanism has to be adjusted to stay effective. The review of the main reactive methods is as follows.

Sniffing and scanning prevention. The literature showcases several approaches to limit the ability of an attacker to sniff traffic for learning the configuration and the properties of the network.[36,37] The common idea is to discover whether a wired/wireless network interface card is set to promiscuous mode, that is, all the received frames are passed to the higher layers of the protocol stack despite the host is not the intended destination. To this aim, two main techniques exist: *challenge-based*[36] and *measurement-based*.[37] In challenge-based methods, the defender provokes a reply from the (supposed) sniffing machine by using ad-hoc crafted network traffic (typically, packets with a forged MAC address). In measurement-based methods, a host suspected to be controlled by the attacker is flooded with suitable traffic patterns. In both cases, the provided answer or its temporal evolution will help the defender to identify the reconnaissance attempt. Alas, the continuous development of hardware and OSs reduces the effectiveness of such techniques, mainly due to the need of having updated templates to compare the received traffic.[9]

Lastly, the advent of automatic and efficient scanning services like Shodan revamped the importance of carefully designing the addressing scheme to be used. In fact, IoT and smart devices could take advantage of IPv6 both in terms of end-to-end transparency and difficulties in performing a brute-force scan to the entire address space. However, IPv6 can directly expose portions of the network, thus the use of private IPv4 schemes jointly with Network Address Translation is a common and early front line defense technique.[28] Nevertheless, classical techniques (like firewalls and

Figure 3. Classification and evolution of the reconnaissance countermeasures.


Human-based Countermeasures
Technology-based Countermeasures
Reactive
Proactive
Reconnaissance Technique
Social Engineering
Internet Intelligence
Network Information Gathering
Side-Channels
Training and Awareness
Public Information Monitoring
Sniffing and Scanning Prevention
Side-Channels Sanitization
Moving Target Defense
Cyber Deception
Time

IDS)[8] should still be considered prime countermeasures against port scanning and fingerprinting attempts. Finally, recent approaches focus also on analyzing backscatter traffic, that is, network traffic generated by unallocated or unused IP addresses in a near real-time manner. Such methods can be used to identify reconnaissance campaigns in industrial control systems scenarios.[10]

Side-channels sanitization. To limit the exposure to side-channels, both hardware and software countermeasures have been proposed to "sanitize" the behaviors responsible for leaking data.[32] Hardware mechanisms include, among others, techniques to limit the signal leakage by utilizing Faraday-cage-like packaging, minimize the number of metal parts of a component, or make the circuitry less power consuming to tame EM emissions. For the case of software countermeasures, we mention tools to randomize the sequences of operation or table lookups as well as mechanisms to avoid specific instructions patterns as to prevent the CPU/GPU radiating distinguishable EM patterns that act as a signature.

Side-channels also let attackers map virtual resources in cloud datacenters or honeypots. Since the cache architecture or the timing behaviors are often abused for this purpose, many countermeasures focus on modifying the underlying OS to introduce time-padding (to assure that execution time of a protected function is independent of any secret data the function operates on), cache cleansing (to forbid obtaining the state of the cache after running the sensitive function), and dynamic partitioning methods (to protect resources of a trusted process from being accessed by an untrusted process during its execution). Other possible countermeasures against side-channels can be deployed within the hypervisor or at the application level.[5]

Lastly, if side-channels are used to infer information through the network, a prime solution is to use some form of traffic normalization. In this case, ambiguities of the flow that can be exploited to infer data are removed by suitable manipulation of Protocol Data Units. For instance, the presence of a firewall can be sensed by inspecting delay and the inter-packet time statistics, thus suitable buffering techniques could prevent to leak such an information.

Side-channels let attackers map virtual resources in cloud datacenters or honeypots.

Proactive technology-based countermeasures. Proactive solutions have been proposed to anticipate attackers by constantly shifting or poisoning the information that can be learned during the reconnaissance phase. The literature reports two main classes of techniques.

Moving Target Defense (MTD)[24] is a recent class of approaches aiming at recovering from the current asymmetry between attackers and defenders. In essence, MTD can limit the exposure of the victim by dynamically varying the configuration of network and nodes in order to make the leaked data unstable or outdated. The price to be paid is in terms of overheads experienced by the defender and legitimate users, for example, delays needed to change configurations and temporary device unavailability due to reassignments of addresses. As a possible example of production-quality MTD mechanisms, Dynamic Network Address Translation[24] allows interfering with malicious scanning phases by replacing TCP/IP header information while assuring service availability. Another method to protect cloud environments is to modify the scheduler to randomly allocate virtual machines and prevent co-residency attacks and side-channels between VMs running on the same physical machine.[5]

Cyber deception. Another emerging proactive cyber defense technology is Cyber Deception (CD). In this case, the defender provides to attackers misleading information in order to deceive them.[40] A possible approach deals with the manipulation of the network traffic to deliver the attacker a virtual, yet useless, network topology.[1] Differently from MTD, a mechanism based on CD does not continuously transform the defended deployment. Rather, it aims at distracting the attacker away from the most critical parts or to route and confine him/her within a honeypot or a honeynet. While a honeypot tries to lure the attacker into a single, deliberately vulnerable system, a honeynet works on a larger scale by "simulating" a whole subnetwork. Thus, observing the attacker operating in such a strictly controlled environment allows to infer indicators of compromise that can be used both for anomaly detection purposes as well as to protect the real network from information gathering attempts.

Proactive countermeasures are expected to evolve into solutions able to combine CD and MTD approaches.[40] In such setups both techniques can be seen as complementary: MTD permits to adapt a system or a network to increase its diversity and complexity, while CD directs adversaries into time-consuming but pointless actions, thus draining their resources.

Conclusion and Outlook

This article has focused on the reconnaissance phase, which is the basis for the totality of cybersecurity attacks.

As a general trend, the evolution of smart devices, social media, and IoT-capable applications, boosted the amount of information that can be gathered by an attacker and also multiplied the communications paths that can be used to reach the victim. Therefore, the potential attack surface exploitable for reconnaissance techniques is expected to continue to grow, at least in the near future.

Regarding the development of countermeasures, defenders appear to be a step back with respect to attackers. To fill such gap, countermeasures should aim to:

- strengthen training and monitoring to also consider threats leveraging side-channels;
- evaluate how to incorporate results obtained via public sources into proactive countermeasures;
- expand solutions exploiting cyber deception also to counterattack social engineering (for example, when an employee detects a scam attempt, he/she intentionally mislead the attacker) and side-channels (for example, by deliberately leaking incorrect information);
- engineer a new-wave of reconnaissance-proof-by design services, for instance, by minimizing the impact of the addressing scheme, the use of IoT and the exposition to scanning services like Shodan; and,
- re-think the concept of privacy in a more broad manner to also include protection mechanisms against advanced and malicious data gathering campaigns.

Acknowledgments
This work has been partially supported by EU Project SIMARGL, Grant Agreement No 833042 and by the Polish National Agency for Academic Exchange (Grant No PPN/BEK/2018/1/00153).

References

1. Achleitner, S., La Porta, T., McDaniel, P., Sugrim, S., Krishnamurthy, S.V., Chadha, R. Cyber deception: Virtual networks to defend insider reconnaissance. In *Proceedings of the 8th ACM CCS Intern. Workshop on Managing Insider Security Threats*, Oct. 2016, 57–68.
2. Al-Saleh,M. Crandall, J.R. Application-level reconnaissance: Timing channel attacks against antivirus software. In *Proceedings of the 4th USENIX Conf. Large-scale Exploits and Emergent Threats*, 2011, 1–8.
3. Aviv, A., Gibson, K., Mossop, E., Blaze, M., Smith, J.M. Smudge attacks on smartphone touch screens. In *Proceedings of the 4th USENIX Conf. on Offensive Technologies*, 2010, 1–7.
4. Backes, M., Dürmuth, M., Gerling, S., Pinkal, M., Sporleder, C. Acoustic side-channel attacks on printers. In *Proceedings of the USENIX Security Symposium*, 2010, 307–322.
5. Bazm, M., M. Lacoste, M., M. Südholt, M. and J. Menaud, J. Side-channels beyond the cloud edge: New isolation threats and solutions. In *Proceedings of the 1st Cyber Security in Networking Conf.*, Oct. 2017, 1–8.
6. Bejtlich, R. The Tao of Network Security Monitoring Beyond Intrusion Detection. Pearson Education, 2004, ISBN: 0-321-24677-2.
7. Berger, Y., Wool, A. Yeredor, A. Dictionary attacks using keyboard acoustic emanations. In *Proceedings of the 13th ACM Conf. Computer and Communications Security*, 2006, 245–254.
8. Bou-Harb, E., Debbabi, M., Assi, C. Cyber scanning: A comprehensive survey. *IEEE Communications Surveys & Tutorials 16*, 3 (3rdQ 2014). 1496–1519.
9. Cabaj, K., Gregorczyk, M., Mazurczyk, W., Nowakowski, P., Żórawski, P. Sniffing detection within the network: Revisiting existing and proposing novel approaches. In *Proceedings of the 5G Network Security Workshop to be held jointly with the 14th Intern. Conf. on Availability, Reliability and Security*, 2019.
10. Cabana, O., Youssef, A.M., Debbabi, M., Lebel, B., Kassouf, M., Agba, B.L. Detecting, fingerprinting and tracking reconnaissance campaignst industrial control systems. *Detection of Intrusions and Malware, and Vulnerability Assessment, LNCS 11543* (June 2019) . R. Perdisci, C. Maurice, G. Giacinto, M. Almgren (Eds.). Springer, 89–108.
11. Caviglione, L., Wendzel, S., Mazurczyk, W. The future of digital forensics: Challenges and the road ahead. *IEEE Security & Privacy 15*, 6, (Nov./Dec. 2017), 12–17.
12. Caviglione, L., Coccoli, M. Privacy problems with Web 2.0. *Computer Fraud & Security 10* (2011), 16–19.
13. Collins, M., Shimeall, T., Faber, S., Janies, J., Weaver, R., Shon, M.D., Kadane, J. Using uncleanliness to predict future botnet addresses. In *Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference*, 2007, 93–104.
14. Genkin, D., Pattani, M., Schuster, R., Tromer, E. Synesthesia: Detecting screen content via remote acoustic side channels. In *Proceedings of the IEEE Symp. Security & Privacy*, 2019
15. Goodman, M. *Future Crimes*. Anchor Books, New York, 2016, ISBN 9780804171458.
16. Holz, T., Gorecki, C., Rieck, K., Freiling, F. Measuring and detecting fast-flux service networks. In *Proceedings of the 15th Network and Distributed System Security Symp.*, 2008, 257–268.
17. Holz, T., Raynal, F. Detecting Honeypots and Other Suspicious Environments. In *Proceedings of the 6th Annual IEEE SMC Information Assurance Workshop*, 2005, 29–36.
18. H2020 Project—Diversity Enhancements for Security Information and Event Management. Project Deliverable D4.1: Techniques and Tools for OSINT-based Threat Analysis; http://disiem-project.eu/wp-content/uploads/2018/06/D4.1v2.pdf
19. *Industry Week*. Cyberattacks skyrocketed in 2018. Are you ready for 2019?; https://www.industryweek.com/technology-and-iiot/cyberattacks-skyrocketed-2018-are-you-ready-2019
20. Kocher, P. Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In *Proceedings of the Annual Intern. Cryptology Conf.* Springer, Berlin, Heidelberg, 1996, 104–113.
21. Krawetz, N. Anti-honeypot technology. *IEEE Security & Privacy 2*, 1 (Jan-Feb 2004), 76–79.
22. Lampson, B. A Note on the confinement problem. *Commun. ACM 16*, 10, (Oct. 1973), 613–615.
23. Lancor, L., Workman, R. Using Google hacking to enhance defense strategies. *ACM SIGCSE Bulletin*, 2007, 491–495.
24. Lei, C., Zhang, H.Q., Tan, J.L., Zhang, Y.C., Liu, X.H. Moving target defense techniques: A survey. *Security and Communication Networks* 2018, 1–25.
25. Li, Z., Yu, X., Wang, D., Liu, Y., Yin, H., He, S. SuperEye: A distributed port scanning system. *Artificial Intelligence and Security LNCS 11635*. X. Sun, Z. Pan, E. Bertino, (Eds). Springer, Cham, July 2019, 46–56.
26. Lockheed Martin. The Cyber Kill Chain; https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html
27. MITRE, ATT&CK Framework; https://attack.mitre.org/
28. Notra, S., Siddiqi, M., Gharakheili, H.H., Sivaraman, V., Boreli, R. An experimental study of security and privacy risks with emerging household appliances. In *Proceedings of the IEEE Conf. on Communications and Network Security*, 2014, 79–84.
29. O'Hare, J., Macfarlane, R., Lo, O. Identifying Vulnerabilities Using Internet-Wide Scanning Data. In *Proceedings of the 12th IEEE Intern. Conference on Global Security, Safety and Sustainability*, pp. 1-10, 2019.
30. Ristenpart, T., Tromer, E., Shacham, H., Savage, S. Hey, you, get off of my cloud: Exploring information leakage in third-party compute clouds. In *Proceedings 16th ACM Conf. Computer and Communications Security*, 2009, 199–212.
31. Salahdine, F. Kaabouch, N. Social engineering attacks: A survey. *Future Internet 11*, 4 (2019), 1–17.
32. Sayakkara, A., N.-A. L.-K., Scanlon, M. A survey of electromagnetic side-channel attacks and discussion on their case-progressing potential for digital forensics. *Digital Investigation 29* (2019), 43–54.
33. Siby, S., Maiti, R.R., Tippenhauer, N.O. IoTScanner: Detecting privacy threats in IoT neighborhoods. In *Proceedings of the 3rd ACM Intern. Workshop on IoT Privacy, Trust, and Security*, 2017, 23–30.
34. Simon, L., Xu, W., Anderson, R. Don't interrupt me while I type: Inferring text entered through gesture typing on Android keyboards. In *Proceedings of Privacy Enhancing Technologies 3* (2016), 136–154.
35. Siponen, M. A Conceptual foundation for organizational information security awareness. *Information Management & Computer Security 8*, 1 (2000), 31–41.
36. Trabelsi, Z. and Rahmani, H. Detection of sniffers in an Ethernet network. *Information Security, LNCC 3225* (Sept. 2004). K. Zhang, Y. Zheng (Eds) Springer, Berlin, Heidelberg, 170–182,
37. Trabelsi, Z., Rahmani, H., Kaouech, K., Frikha,M. Malicious sniffing systems detection platform. In *Proceedings of the Intern. Symp.Applications and the Internet*, 2004, 201–207.
38. Trestian, I., Ranjan, S., Kuzmanovic, A., Nucci, A. Googling the Internet: Profiling Internet endpoints via the World Wide Web. *IEEE/ACM Trans. Networking 18*, 2 (2010), 666–679.
39. Wang, P., Wu, L., Cunningham, R., Zou, C.C. Honeypot detection in advanced botnet attacks. *Intern. J. Information and Computer Security 4*, 1 (2010), 30–51.
40. Wang, C., Lu, Z. Cyber deception: Overview and the road ahead. *IEEE Security & Privacy 16*, 2 (M-A 2018), 80–85.

Wojciech Mazurczyk is University Professor at Warsaw University of Technology, Institute of Computer Science, Warsaw, Poland.

Luca Caviglione is a senior research scientist at National Research Council of Italy, Institute for Applied Mathematics and Information Technologies, Genova, Italy.