Chapter XVI
Network-Based Passive Information Gathering

Romuald Thion
University of Lyon, France

ABSTRACT

Please submit an abstract for your chapter (a brief introduction of the chapter and/or explanation of what topics will be covered in the chapter). Your abstract should be between 150 and 200 words in length.

INTRODUCTION

The rise of the Internet has been a blessing for computer science and the world of economy. It has redefined the word "information"; the Internet is the tip of the information revolution iceberg. The information revolution implies the rise of a mode of warfare in which neither mass nor mobility will decide outcomes; it is the new concept of "cyber war." It means trying to know everything about an adversary via network interconnections, while keeping the adversary from knowing much about him or herself. This tactical principle has already been exposed by Tzu in his *Art of War* (1910), but clearly it takes a new dimension in our interconnected world:

what enables the wise sovereign and the good general to strike and conquer, and achieve things beyond the reach of ordinary men, is foreknowledge. Now this foreknowledge cannot be elicited from spirits; it can-

not be obtained inductively from experience, nor by any deductive calculation. Knowledge of the enemy's dispositions can only be obtained from other men.

In this topic, we are specifically interested in *passive* network-based information gathering techniques. In the context of networks, passive refers to techniques that do not connect to the targeted system or that would not be normally associated to an attack, whereas *active* refers to techniques that create network traffic and could be associated with suspicious or malicious behavior (e.g., port scanning).

BACKGROUND

Penetration testers, ethical hackers, and cyber criminals conduct cyber attacks in the same way. Whereas penetration testers are reliable people paid by an organization to conduct a security audit by "attacking" the target to find vulnerabilities and security weaknesses, cyber criminals and "nonethical" hackers conduct attacks without an organization's consent, to earn money, to undermine the credibility of the target, or for any other motive. In both cases, the techniques are identical. An attack can be roughly separated into five steps (FX et al., 2004).

1. **Information gathering:** By gathering as much information as possible about the target, in this step, the hacker is looking for potential vulnerabilities as well as software and hardware in use, network topology, and any information that will be useful for its attack (Grubb, 2004).
2. **Exploitation:** Using foreknowledge, a cyber criminal can focus on a specific vulnerability to take the initiative. In this step, the hacker is trying to find the most powerful and least difficult way to exploit vulnerability.
3. **Privileges elevation:** Often an exploited vulnerability does not award full control of the system. In this step, the hacker elevates his privileges to root around any means available.
4. **Cover tracks:** Once a system has been compromised, the hacker wants to cover his tracks as soon as possible, thus providing more time to act and lessen the possibility of being caught.
5. **Carry out his objective:** The hacker reaps the fruits of his or her efforts. He or she can gather any sensitive information wanted, use the compromised system to attack another one, delete data, and so forth. The hacker achieves the attack objectives.

This topic focuses on the very first step of any attack, information gathering, also known as preassessment information gathering. During this phase, the attacker is interested in obtaining preliminary information about the target—the foreknowledge.

Information gathering techniques can be roughly classified into the following:

- **Social engineering:** These nonnetwork-based techniques are the practice of obtaining confidential information by manipulating users. A social engineer fools people (e.g., by phone, by e-mail) into revealing sensitive information or getting them to do something that is against typical policies, using their gullibility. Social engineering is made possible because "the weakest link of the security chain is the human factor." For instance, the famous hacker Kevin Mitnick has extensively used these techniques (Mitnick, Simon, & Wozniak, 2002).
- **Active:** This includes intrusive reconnaissance that sends (specially crafted) packets to the targeted system, for example, port-scanning. Advanced network enumeration techniques avoid direct communication with the targeted host (e.g., Nmap (Fyodor, 2006)).
- **Passive:** This includes reconnaissance that either does not communicate directly to the targeted system or that uses commonly available public information, not normally identifiable from standard log analysis (Zalewski, 2005). This topic is focused on this category.

Every Internet-connected system unintentionally leaks internal information about its organization, making the passive information gathering process possible. Moreover, many organizations fail to identify potential threats from information leakage that could be used to build an attack.

Most information about an organization is publicly available using the Internet, or contained on systems unrelated to the target. This kind of information can be accessed anonymously by anyone without ever coming into direct contact with the organization's system; this is an important aspect of information leakage. Passive information gathering techniques could be applied to any public service available, for instance, job announcement services, public reports, or public directories.

PASSIVE TECHNIQUES

This section reviews traditional network-based passive information gathering techniques. All of these techniques use unsuspicious-looking connections. Most of them are based on collecting and harvesting publicly available information, such as:

- **"Real-life" information**, for example, physical locations, real names, telephone numbers, the internal structure of organizations, business processes, and so forth that could be later used in social engineering techniques (by endorsing an employee's identity, for example, Mitnick et al., 2004). This kind of information broadens an attacker's knowledge of the victim, making his attack well-targeted.
- **"Technical" information**, for example Internet protocol (IP) addresses, network topology, and software and hardware versions of both servers and clients. This kind of information helps the attacker to find the weakest link of the security chain. An attacker's goal cannot be reached directly, most of the time, instead the attacker needs to breach the systems by using the most simple and effective way. Technical information can reveal easy to exploit vulnerabilities or interesting devices he or she needs to control to reach his goal.

Internet Service Registration

Every accessible host over the Internet must have a unique IP address (e.g., 207.46.20.60). To simplify host addressing and its usage by human beings, the domain name system (DNS) associates IP addresses to a unique domain name (e.g., microsoft.com).

International structures manage both IP addresses and domain names. Organizations must supply administrative information to these international instances, which is publicly available and may be accessed freely by anyone. Querying those international databases is the very first step in information gathering. The whois resource service provides a mechanism for querying these databases. Among the useful information provided are physical location, real names, and phone numbers. This information is particularly useful for social engineering (Ollman, 2004). The dNS addresses, the number of IP addresses attributed, Internet service provider (ISP) contact and registrar can reveal sensitive technical information. Table 1 is an extract of a whois query revealing phone numbers, physical addresses, and real names. A collection of tools related to domain name and IP registration can be found at http://www.dnsstuff.com/.

Domain Name Service

Most operating systems (OS) include the *name service look up* (nslookup) tool (Mockapetris, 1987). The Unix based OS includes the *dig* tool as well. These tools are made to query DNS records (on DNS service, such as BIND, TCP/UDP port 53). They can provide extensive valuable information to an attacker. They can be used to resolve names into IP addresses and vice versa. One of the most powerful functionalities is the "zone transfer," where complete DNS records are transferred from one DNS server to another, but it can be manually executed using nslookup or dig, thus providing exhaustive information about the targeted

Table 1. A sample name service-based whois result: "whois -h whois.nic.fr univ-lyon1.fr"

```
domain:      univ-lyon1.fr
address:       Centre Informatique Scientifique et Medical de
l'Universite Claude Bernard Lyon 1
address:     batiment 101, 27 A 43 boulevard du 11 Novembre 1918
address:     69622 Villeurbanne Cedex
address:     FR
phone:       +33 4 72 44 83 60
fax-no:      +33 4 72 44 84 10
e-mail:      gilles.rech@univ-lyon1.fr
admin-c:     GR258-FRNIC
tech-c:      GR1378-FRNIC
zone-c:      NFC1-FRNIC
nserver:     dns.univ-lyon1.fr 134.214.100.6
nserver:     dns2.univ-lyon1.fr 134.214.100.245
nserver:     ccpntc3.in2p3.fr 134.158.69.191
nserver:     ccpnvx.in2p3.fr 134.158.69.104 …
```

Table 2. Sample reverse (from IP to name) DNS results

```
smtphost.example.com(192.168.0.4), mail server
dns.example.com(192.168.0.6), dns server
pop.example.com(192.168.0.7), mail server
routeur-ipv6-v100.example.com(192.168.0.45) , IPV6 router
dhcprov100-02.example.com(192.168.0.47), DHCP server
testmath.example.com(192.168.0.231), promising "unsecure" host
cisco-ls.example.com (192.168.4.9), cisco router
hpserv. example.com (192.168.4.10), Hewlett-Packard server
```

organization (Barr, 1996). The interesting information includes e-mail servers names (and addresses), Web servers, routers and firewall addresses. Most of the time, sensitive information can be deduced from the organizational naming convention, such as software and hardware information (e.g., OS, constructor), services available, and so forth (Grubb, 2004). For instance, some illustrative results are shown in Table 2 for example.com, which is a registered domain name.

E-Mail Systems

If Web sites provide the shop front of business organizations, e-mail provides essential business communication systems. A lot of information can be collected through the analysis of mail systems. Simple mail transfer protocol (SMTP) (Postel, 1982) is the standard protocol for e-mails. The analysis of its header can provide internal server naming, topology of network, user accounts, a version of e-mail services, clients, patch level, type and version of content filter, and antispam or antivirus solutions. Table 3 shows a sample SMTP header. It can be seen that this e-mail was sent by "Sample User" whose address is user@example.com, using Microsoft Outlook on a laptop with the IP address 192.168.5.26. This sample does not include the SMTP relay, but analyzing the chain also is very useful. It can reveal trusted relationships between e-mail servers and internal topology. According to this example in which Outlook 2000 (Microsoft Outlook Build 9) is used by Sample User, a cyber attacker may

Table 3. Sample SMTP header

```
Return-Path: <user@example.com>
Received: from cri14.sample.fr
     by dsi02.sample.fr (Cyrus v2.2.12) with LMTPA;
     Wed, 22 Feb 2006 12:02:37 +0100
…
Received: from out4.example.fr
    by cismrelais. sample.fr (Postfix) with ESMTP id 8417E48104
    for <john.doe@dummy.com>; Wed, 22 Feb 2006 11:52:20 +0100
Received: from UserLaptop ([192.168.5.26]) by out4.example.fr
(Sun Java System Messaging Server 6.1 HotFix 0.11 (built Jan 28 05))
Date: Wed, 22 Feb 2006 12:51:58 +0200
From: Sample User < user@example.com >
Subject: Sample test, France
In-reply-to: <34f699a5f6e6a879072a609ea2b46d6d@example.com >
To: "'John DOE'" <john.doe@dummy.com >
X-MIMEOLE: Produced By Microsoft MimeOLE V6.00.2800.1106
X-Mailer: Microsoft Outlook CWS, Build 9.0.2416 (9.0.2910.0)
X-Virus-Scanned: by AMaViS snapshot-20020222
X-Virus-Scanned: amavisd-new
```

focus on Outlook vulnerabilities to break into example.com, or he may try to exploit Microsoft Office 2000, conjecturing that it is used by the company.

Web Site Analysis

The larger or more complex a Web site is, the higher the probability of it inadvertently leaking internal information, and the more information an attacker can obtain. Large sites can be managed by several administrators, built by dozens of developers, and filled in by hundreds of people; this may lead to information disclosure. A common technique for an attacker is to retrieve the whole targeted site and to analyze its content on his local image, thus avoiding multiple suspicious connections. The hackers will freely explore and harvest the site for sensitive information. The process of automatically retrieving a Web site and analyzing its content is commonly referred to as "Web crawling." Common tools for Web scraping are Sam Spade and Wget. Sam Spade crawls and discovers linked Web pages on a site. This is an efficient tool that can quickly download a company's entire Web site. Another very powerful tool is Wget, a scriptable command-line browser. It can grab HTML pages, images, and forms as a "standard" browser.

Interesting findings include (Ollman, 2004):

- Real names and e-mail addresses
- Comments from internal developers can reveal technical information about technologies in use, maintenance operations, internal resources, or connectivity methods (e.g., database connector). Badly cleaned sources can even reveal pieces of server-side code or even default passwords.
- Comments can reveal debug, prototype, or test information, such as disabled pages or internal development hosts that would be normally inaccessible.
- Signature of tools (within metatags, for example) can give very precise information about version and development software.
- Logs and temporary files are very fruitful findings that can reveal very sensitive details, like user habits or links to external customer Web sites.
- Error pages, such as 404 (page not found) and 500(internal error), can be fruitfully exploited. They can reveal the existence (or absence) of files, coding errors, or dead URLs.
- Links to documents and binary data may suffer from great leakage. For example, Microsoft Word

files usually include internal host names, real names, and even shared resource locations.

Thus, it is very important that all content be analyzed and cleaned for any unintentional leakage.

CURRENT ISSUES

Techniques discussed in the previous section are based on publicly available information; domain registration, DNS, mail headers. Web content, and binary data available over the Internet also were discussed. Whereas the first ones imply the use of dedicated (although very common) tools, such as dig, whois, or *traceroute*, there exists an extremely powerful tool that crawls the Internet with very accurate and efficient querying capabilities of Web content—the Google search engine (Long, Skoudis, & Van Eijkelenborg, 2001).

Google's cache system, advanced query operators, such as *site:*, *filetype:*, *intitle:*, or even translation services, makes it a major tool in the passive information gathering arsenal. We will describe a few techniques using Google that can be successfully applied to gather information without any direct connection to the target and to harvest Web content that should be kept private.

- **Using the cache system:** Google keeps snapshots of crawled pages in its own repository. You may have experienced it using the "cached" link appearing on search results pages. The advanced operator cache: is used to jump directly to the cached snapshot of a Web site without performing a query. This is a quite simple and effective way to browse Web pages without any direct connection to the target.
- **Using Google as a proxy server:** Google can be used as a transparent proxy server via its translation service. When you click on the "translate this page" link, you will be taken to a version of the page that has been automatically translated into your language. You can use this functionality to translate a page into the same language it is written in, thus, Google crawls the page, does nothing in the translation process (e.g., from English to English) and gives you back that page. This trick can be done by modifying the hl variable in Google search URL to match the native language of the page.
- **Discovering network resources:** Google can help with the network discovery phase. Google searches can be seen as an alternative to DNS queries, by combining the site: operator and logical NOT, a hacker can obtain a list of public servers. For example, "site:microsoft.com -www.microsoft.com" will reveal msdn.microsoft.com, directory.microsoft.com, partnerconnect.microsoft.com, officelive.microsoft.com, and so forth. Moreover the *link:* operator finds pages that link to the queried URL; it can be used to provide important clues about relationships between domains and organizations. The *intitle:* and *inurl:* operators can be used to detect the presence of Web-enabled network devices, such as routers. For example, *inurl:tech-support inurl:show Cisco OR intitle:"switch home page" site:example.com* searches Cisco's Web-enabled devices on the domain example.com.
- **Harvesting system files, configuration files, and interesting data using advanced specific queries:** Hundreds of Google searches can be found in Long et al. (2001). Their book describes in depth advanced operators and how to use them to find passwords (clear or hashed), user names, Web-enabled devices, and so on. Table 4 presents simple, but powerful, Google searches that can be processed to retrieve system files, configuration files, and specific data. The main idea is to combine operators, such as *intitle:*, *inurl:*, and *site:*, with specific sentences. For example "# -FrontPage-" is a banner from FrontPage files. The 10 queries in Table 4 are realistic sample queries that can be successfully processed to find passwords or configuration files.

Table 4. Ten security queries that work from johnny.ihackstuff.com

```
1) "http://*:*@www" domainname (get inline passwords)
2) intitle:index.of.password (or passwd or passwd)
3) "access denied for user" "using password" (SQL error message, this message
can display the username, database, path names and partial SQL code)
4) "AutoCreate=TRUE password=*" (Searches the password for "website access.
Analyzer")
5) intitle:"Index of" _vti_inf.html ("vti_" files are part of the FrontPage
communication system between a web site and the server)
6) "# -FrontPage-" ext:pwd inurl:(service | authors | administrators | users) "#
-FrontPage-" inurl:service.pwd (search for MD5 hashed FrontPage password)
7) inurl:passlist.txt
8) "A syntax error has occurred" filetype:ihtml (Informix error message,
this message can display path names, function names, filenames and partial
code)
9) allinurl:auth_user_file.txt (DCForum's password file. This file gives a
list of passwords, usernames and email addresses)
10) allinurl: admin mdb (administrator's access databases containing user-
names, passwords and other sensitive information)
```

CONCLUSION

Most organizations and system administrators are familiar with penetration-testing and intrusion-detection techniques. These techniques are cornerstones in security evaluation and focus mainly on the exploitation of vulnerabilities and suspicious/malicious behavior (e.g., log analysis). However, an organization relying mainly on these techniques may underestimate the huge amount of information that can be anonymously obtained from publicly available content over the Internet. This topic gives an overview of network-based passive information gathering techniques. Some can note that passive techniques are also very useful from an internal perspective; it reduces traffic within the internal network (e.g., passive OS fingerprinting to enumerate OSs in use (Treurniet, 2004)). To protect themselves, organizations should carefully check their publicly available information.

- Some information must be published (e.g., contact e-mail), but protection measures should be established to prevent automated crawlers from finding this information, if it can be misused (e.g., for spam). A common way to avoid sensitive information being crawled is to protect it by mechanisms simple for humans but complex for machines. For example, regular expressions cannot match "[at]" images within e-mail addresses (do not write e-mail clearly).
- The principle of the least privilege must be respected by publishing only a strict minimum, denying bots the ability to crawl public but sensitive information. This advice is legitimate for DNS; do not publish names of hosts or devices that should not be accessed from the outside Internet. It is also legitimate for configuration files. If a file is not meant to be public (e.g., _vti_ files for FrontPage, debug/test pages), keep it private.
- Conduct reviews of code and Web pages to keep them clean and avoid comments, prolix banners, version numbers, and so forth . A lot of information can be gathered from error pages, banners, and seemingly innocuous information. Comments can be incredibly information leaking; entire blocks of server side code within client's pages are not so uncommon.

To sum up, information gathering is the very first step of an attack and probably the most crucial in achieving the attacker's goal. Information collected

in this phase is raw material that is used to build a firm attack. The attackers can obtain a global view of the target, can focus on the weakest link in security, and can obtain enough information to conduct social engineering. If conducted cleanly via passive techniques using publicly available information, this step is anonymous and practically undetectable. Thus, organizations should be very careful with content anonymously available over the Internet and should take simple, but effective, measures.

Your physical mail box should be accessible to anyone, at least your mailman. However, nobody will write his own Social Security number, birth date, or job on his or her mail box in the real world. Such information must be kept private from mailmen and passers-by; it should be the same in the cyber world.

REFERENCES

Barr, D. (1996). RFC 1912: Common DNS operational and configuration errors.

FX, Craig, P., Grand, J., Mullen, T., Fyodor, Russell, R., & Beale, J. (2004). *Stealing the network: How to own a continent*.

Fyodor. (2006). *Nmap (network mapper) documentation (including zombie scanning technique).* Retrieved from http://www.insecure.org/nmap/docs.html

Grubb, L. (2004). *Survey of network weapons: part 1: weapons for profiling.* Consortium for Computing Sciences in Colleges (CCSC).

Long, J., Skoudis, E., & Van Eijkelenborg, A. (Eds.). (2001). *Google hacking for penetration testers*..

Mitnick, K., Simon, W., & Wozniak S. (2002). *The art of deception: controlling the human element of security.*

Mockapetris, P. (1987). RFC 1035: Domain names—Implementation and specification.

Ollman, G. (2004). *Passive information gathering: The analysis of leaked network security information* (Tech. paper). Next Generation Security Software Ltd.

Postel, J. (1982). *RFC 821: Simple mail transfer protocol.*

Treurniet, J. (2004). *An overview of passive information gathering techniques for network security* (Tech. memo.). Defence R&D Canada.

Tzu, S. (1910). *Sun Tzu on the art of war, the oldest military treatise in the world*. (L. Giles, Trans.).

Zalewski, M. (2005). *Silence on the wire: A field guide to passive reconnaissance and indirect attacks.*

TERMS AND DEFINITIONS

Domain name system or domain name server or domain name service (DNS): This is a system that stores information associated to domain names. The most important being the Internet protocol (IP) addresses associated with a domain name, but it also lists mail servers and administrative contacts. The domain name system makes it possible to attach a "hard-to-remember" IP address (e.g., 66.249.93.99) to an "easy-to-remember" domain name (e.g., google.com).

Proxy server: This computer offers a computer network service, allowing clients to make indirect network connections to other network services. It acts as a relay of service, including filtering and caching capabilities (e.g., Web proxy that denies access to black-listed sites). A client connects to the proxy server and requests a connection; the proxy provides the resource either by connecting to the specified server or by serving it from a cache.

Simple mail transfer protocol (SMTP): This is the de facto standard for e-mail transmission across the Internet. SMTP is a simple text-based protocol (SMTP commands are commonly achieved by telnet for test purpose), using TCP port 25. To determine the SMTP server for a given domain name, the mail exchange (MX) DNS record is used.

Social engineering: This is the practice of obtaining confidential information by manipulation of legitimate people. Social engineering is used by hackers (e.g., Kevin Mitnick, a famous social engineer) as an effective technique to achieve their goal. It is agreed that exploiting computer vulnerability is often more difficult than tricking people. In order to enhance his or her credibility against the target and to build up trust, a social engineer needs accurate, truthful, and convincing information.

Web crawling: A Web crawler (or Web spider) is a program that browses Web pages in a automated manner. Crawling the Web enables the creation of a copy of all visited pages for later processing, via a search engine, for example. Web crawling permits gathering specific information, such as e-mail (usually for spam).

Whois: This is a query/response protocol that is used for determining owners of domain names and IP addresses or autonomous system information. This system originates as "white pages" for system administrators to contact their peers. Nowadays, it is used to find certificate authority of secured Web pages. Data returned from a query can be used by hackers to broaden their knowledge of a system or for spam (e.g., bot automatically processing whois records to build e-mail bases).

Zone transfer: It is a type of DNS transaction used to replicate DNS databases across DNS servers. The opcodes (used in the "dig" tool, for example) associated with this type of transaction are AXFR (full transfer) and IXFD (incremental transfer). Zone transfer is a way for hackers to manually obtain content of a zone.