THE VOID – AN INTERESTING PLACE FOR NETWORK SECURITY MONITORING

Alexandre Dulaunoy, Gérard Wagener

CIRCL-Computer Incident Response Center
41, av. de la Gare, L-1611 Luxembourg
Tel: +352 24788444
{alexandre.dulaunoy, gerard.wagener}@circl.lu

Marc Stiefer, Cynthia Wagner

Fondation RESTENA, CSIRT
6, rue R. Coudenhove-Kalergi, L-1359 Luxembourg
Tel: +352 4244091
{marc.stiefer, cynthia.wagner}@restena.lu

Paper type

Research paper

Keywords

IP-darkspace monitoring, traffic analysis, sensitive data, security and privacy issues, spelling mistakes

Abstract

The Internet void under normal condition is a boring place because it should be free of any content. By deeply inspecting a "blackhole" monitoring dataset – a dataset from unused IP-address space, surprising and frightening observations can be made. Questions that arise here are: why is there traffic, is it there on purpose or coincidence and does it feature any security impacts. In this article, we highlight some examples of our journey to the noise of the Internet which range from badly configured systems to various unexplained events and consequently to leaked private data. Furthermore, we analyse the effects of spelling mistakes in network data and present some hints to prevent potential undesired consequences.

1. Introduction

With arising new threats in the Internet, the monitoring of unused IP-address space has become a precious source of information to support early warning and security systems in networks. In general, networks on enterprise level hold some sets of IP-address space that is not in use, but globally announced and monitored. These network parts are called darkspaces or blackholes. Any traffic arriving on these IP-address sets is not meant to arrive there, since there are no active hosts on these network parts. This unidirectional traffic is then called noise. Parts of this noise traffic reaching the darkspace is data originated from attacks such as scanning traffic and malware, whereas other parts are due to bad configurations of devices.

Another reason why there is traffic is a human issue. Humans are vulnerable to spelling mistakes, may this be in the simple writing of a manuscript, typing a text on a computer or while configuring complex network devices. Independent of a task, a spelling mistake may provoke undesired side-effects. In network management, these human errors lead to wrong configurations with some strange consequences. Addressing wrong network parts due to a spelling mistake (while hitting the wrong key when typing an IP-address), can lead to the undesired effect of disclosing sensitive information to public, such as passwords of routers or other network devices.

In this article we present our work in progress and present some observations from our operated darkspace sensor. It will be specifically focused on the analysis of wrong configurations due to spelling mistakes by humans. Recommendations will be given from the observations made in the experiments.
This paper is organized as follows: Chapter 2 introduces to blackhole monitoring and the occurring threats. Chapter 3 presents the methodology and Chapter 4 describes the observations and evaluations of the experiments. Other work relevant for this topic will be presented in Chapter 5. Chapter 6 will draw conclusions and present future work.

2. Darkspace, blackhole monitoring and spelling mistakes

In 2005, Bailey et al. [Bailey05-2] define blackhole monitoring as the monitoring of dark address space. A blackhole can be described as routable non-used IP-address space on a computer network. Traffic arriving on this network part is unidirectional and unsolicited. It is announced globally on the Internet and visible to the majority of the Autonomous System[1] (AS) just like any other used address space. Today, IPv4 address space has become a scarce resource and the definition of blackhole monitoring been extended to the monitoring of a temporarily unused address space, called dynamic darkspace.

The task of blackhole monitoring can be described as sampled measurements of the Internet noise and the used collected information represents only a subset of this Internet noise. In general, the observed traffic behaviour can be described to be erratic as it not only depends of the size of the monitored blackhole subnet space, but also on the events occurring on the Internet as a whole. Common with other work on darkspace analysis [Zseby12, Zseby13, Wustrow10…], the observed traffic can be mostly attributed to worm and botnet respectively other attack tool activities such as scanning or probing. Other traffic observations for the blackhole include the side-effects from malware activities, such as Backscatter traffic that can be described as traffic from legitimate hosts under attack.

Another source for unwanted traffic in a blackhole is misconfiguration. By analysing strange traffic on the blackhole, it can be observed that there are two categories of misconfigurations. One category can be described as the configuration errors of devices with default values (e.g. badly copied values from tutorials) and the second category can be described as traffic holding spelling mistakes for legitimate configurations of devices.

This article will focus on the second category of misconfigurations, the analysis of spelling errors in blackhole traffic. Research in Linguistics [Pol83, Kuk92] has shown that there are different categories of spelling mistakes in typing, which will be explained in the following section. The more, the IP-darkspace used in the experiments is well-suited for analysing spelling errors in device configurations in private network address space (RFC1918) and it will be shown that spelling errors also play a significant role in the collected data from our dark address space.

3. Methodology

3.1 Definition of a word

Referring to the Oxford dictionaries[2] in speech and writing, a 'word' can be defined as '*a single distinct element of speech/writing, used (sometimes alone or) with others to form a sentence…'*. In the framework of analysing traffic from a monitored blackhole the definition of 'word' has to be adapted.

In this context a *word* can be defined as an *IP-address* since a part from the extracted traffic of the monitored blackhole is IP-relevant information. Respecting RFC1518 that introduces the CIDR[3] format, a word respectively an IP-address is composed of a network part and a host identifier part, e.g. the IPv4-address 192.168.0.0/16 can be explained by 192.168.0.0 as the network part and /16 the routing prefix size. But in this paper, we consider an IP-address as a string of digits and dots. This definition corresponds to the dotted decimal notation and is frequently used in input forms or configuration files.

The definition of a word can be extended by describing a word as a sequence of digits and dots with a maximum length of 15 characters and a minimum length of 7. The format of the IP-address can be restricted to numbers ranging between 0 to 255 (e.g.: [0-255].[0-255].[0-255].[0-255]) and each digit in a number represents a position in the word. For convenience the positioning is done from left to right. An example of a word is given in Figure 1.

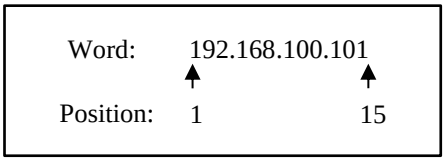

Figure 1: Example of a word with its positions

[1] An AS is a unit of a routing policy or a collection of links and routes for an operator and is based on the Border Gateway Protocol (BGP). BGP is a path vector protocol on address prefix level to achieve high end-to-end connectivity in the Internet.
[2] Definition of word by Oxford dictionaries: http://www.oxforddictionaries.com/definition/english/word
[3] CIDR: Classless Inter-Domain Routing – a Standard scheme for IP-address allocation and IP packet routing.

3.2 **Spelling mistakes**

3.2.1 Spelling mistakes in typing

Spelling mistakes are since ever a well-known failure of human while typing or writing, independent of the category of text. Referring to [Kuk92, Poll83] in text processing it can be distinguished between four basic error types: insertion, omission/deletion, substitution and transposition.

In [Gates37, Poll83 and Kuk92] the different types of errors are defined as follows:

- Insertion – In this case a needless character is accidentally added to the complete word or a character is doubled in the word. Examples: '*ex**x**ample', 'ex**y**ample'...*
- Omission/Deletion – Here, a character is removed from a word or simply forgotten while typing. Examples*: 'exmple', 'helo'...*
- Transposition – In this case, two adjacent characters in a word are interchanged. Examples: *'exam**lp**e', 'netw**ro**k'…*
- Substitution – Here, a character in a word is replaced by another character. This phenomenon happens quite often when typing and hitting the wrong keyboard key. Examples: *'examp**k**e', 'h**w**llo'…*

In this paper we discuss these categories of errors applied to darkspace traffic in order to analyse the purpose of traffic arriving there. Even if in literature [Gates37, Kuk92, Poll83] many more types of errors are used, such as phonetic errors or the word-length phenomenon, we do not consider them relevant for this article.

3.2.2 Position of spelling mistake(s) in a word

An interesting characteristic of spelling mistakes is the position of a mistake in a word. Referring to research of [Gates37, Kuk92, Poll83], errors can occur on different places in a word, but [Kuk92] concluded that 94% of all errors in a word are single errors and belong to the previously cited categories. The most occurring category of mistakes in a text is "omission" reaching about 34%.

Spelling mistakes may occur on each position in a word, but research showed that there are more likely positions for errors to occur in a word. A relevant fact by [Kuk92] is that 23% of all errors in word occur on the 3rd character of a word. In this article, we want to validate the observations from general writing to the IP-addresses observed in the blackhole. Table 1 illustrates the error scheme with the maximum number of possible positions where errors in a generic IP-address, of format 255.255.255.255, for all categories can occur. When applying this error-scheme onto a specific IP-address some restrictions occur to meet the monitoring task since the length of the IP-address may vary. The more, positions with the separator '.' are considered to be error-free. Input forms usually just permit to modify the digits or the program return an error when another character than '.' is specified.

Position	Insertion	Omission	Transposition	Substitution
1	Yes	Yes	Yes	Yes
2	Yes	Yes	Yes	Yes
3	Yes	Yes	Yes	Yes
4	.	.	.	.
5	Yes	Yes	Yes	Yes
6	Yes	Yes	Yes	Yes
7	Yes	Yes	Yes	Yes
8	.	.	.	.
9	Yes	Yes	Yes	Yes
10	Yes	Yes	Yes	Yes
11	Yes	Yes	Yes	Yes
12	.	.	.	.
13	Yes	Yes	Yes	Yes
15	Yes	Yes	Yes	Yes
15	Yes	Yes	Yes	Yes

Table 1: Error-scheme - Example of possible error-positions in an IP-address

3.3 **Project implementation**

The following section presents the different modules used to operate the darkspace probe, to extract and analyse the data. Figure 2 illustrates the implementation of a darkspace sensor in a network and shows, how traffic is routed to it. In Figure 3, an overview of the implementation for a data collector on a darkspace with its related programs is presented.

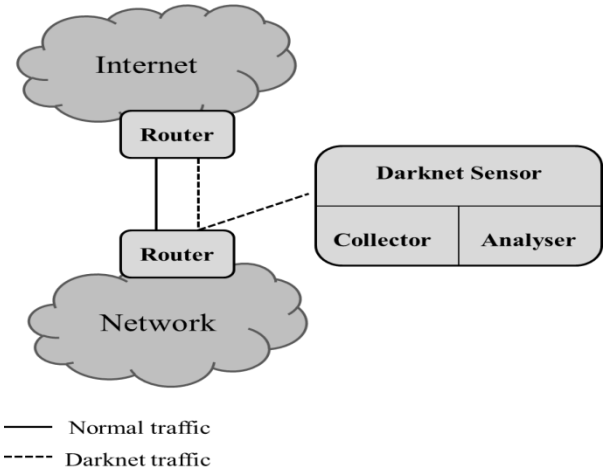

Figure 2: Darkspace architecture

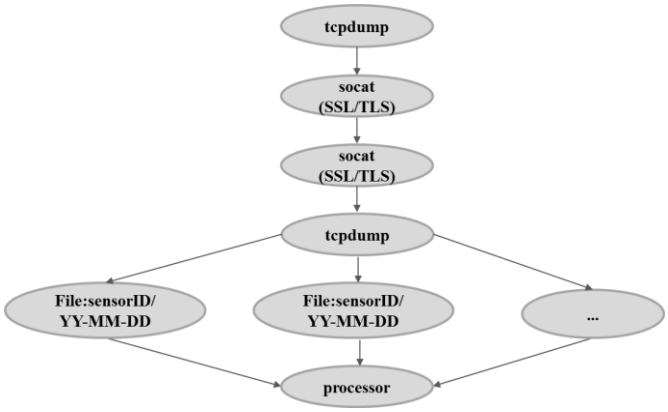

Figure 3: Sensor for data collection

3.3.1 Data collection

The blackhole data is collected from sensors capturing Ethernet frames on a dedicated network part where traffic is routed to. While designing a long-term collection mechanism for blackhole monitoring, the erratic behaviour has to be taken into consideration to cope with low and high bandwidth usage (e.g. it is not unusual to see bandwidth usage variations of a factor ten), as the traffic behaviour may deviate at regular intervals.

Figure 3 shows the process and interaction between the different programs to perform the task of the data collection and its processing. The Ethernet frames are captured on the sensor with tcpdump[4] which writes the raw packets including the pcap-header (containing timestamps and packet lengths) to a standard output. The output of tcpdump is then read by the program socat that establishes an SSL/TLS connection with the collector that also runs an instance of the socat[5] program. This program then ships the Ethernet frames to tcpdump and writes them in a current file. As shown on Figure 3, each file has a filename composed of the sensor name with its date.

The entire data collection chain is tested with the Netbeacon [Netbeacon] program that sends a beacon at regular intervals to a set of monitored addresses. This beacon is transmitted over UDP and includes a timestamp, a sequence number, and a hmac[6] (Hash-based message authentication code) to ensure the integrity of the packet. The processor inspects the recently created files (5 minutes old) and checks the netbeacons. If there is none, the SSL/tunnel probably collapsed and an intervention is needed. The timestamp within a netbeacon is also compared to the timestamp generated by the first tcpdump instance. A large time difference gives an indication about a non-functioning time synchronization service. The processor has to reorganize the files per day in one directory.

A compressed file is processed with different tools to shape its information. Then they are aggregated into indexed documents for further queries and analysis. By this, countries and organizations, which own ASNs, can be ranked by their attacks/misconfigured systems.

[4] http://www.tcpdump.org/
[5] http://freecode.com/projects/socat
[6] http://tools.ietf.org/html/rfc2104

3.3.2 Dataset description

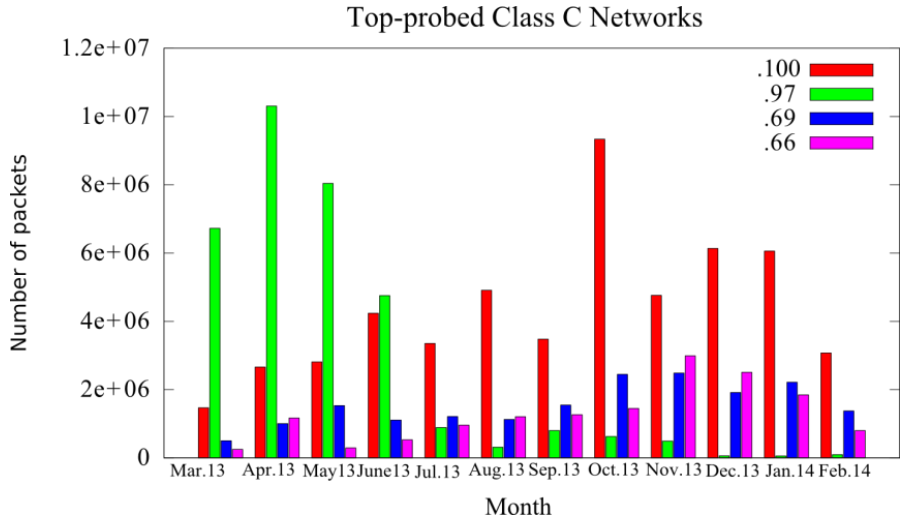

Figure 4: The most probed subnets in the blackhole

In the evaluation part of this paper, only a subset of the overall dataset is analysed. This dataset covers the time period from 2013-03-12 to 2014-02-18 and consumes up to 71 GB of disk-space. The dataset is a set of compressed PCAP files. The monitored blackhole is especially well-suited for spelling mistake detection related to private networks address space since; the blackhole network part is very similar. The real blackhole IP-address ranges will not be disclosed in the paper due to confidentiality issues, but it spawns over multiple /24 networks and will use the anonymized IP-address xyz.xyz.[0-255].[0-255] in the whole article.

In Figure 4 the most probed subnets in the xzy.xyz.0.0/24 are represented. It can be observed that the xyz.xyz.100.0/24 is more likely to be probed than the other subnets. We assume that the popularity of the .100 network is because it is easy to remember and this subnet is often used in literature as a reference for a home network. The .69 and .66 are less popular but their presence is more or less stable over time. However, the network .97 seems to be only temporarily popular and may represent a temporarily misconfigured device.

Figure 5 gives a general overview of the DNS data collected between June and August 2013. The analysis mainly focuses on DNS traffic in this article, but also general evidence extracted from the overall monitored traffic will be included to extend the completeness in the security recommendation part.

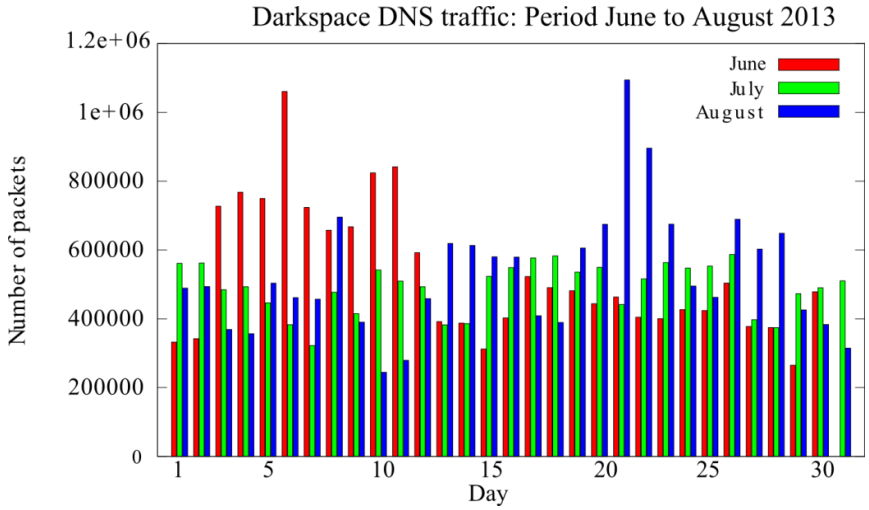

Figure 5: Darkspace DNS Traffic pattern for 3 months in 2013

4. Results and Interpretation

4.1 General observations

A lot of packets arriving on this network relate to scanning, attacks, malware (backscatter traffic, conficker, etc.) or are accidental misconfigurations of devices that leak information. On a first sight, the case of badly configured network devices might seem harmless, because a logical consequence of a wrongly configured service may be that it is unavailable, but in this article we show that also significant information can be leaked.

In general, it can be said that the monitored traffic on the darkspace probe follows a general network traffic pattern as for normal operational networks. Figure 5 represents the monitored traffic for a period of three months. It can be observed that there is more traffic monitored on week-days than on week-ends. For example, June 1st and 2nd was a week-end, here, it clearly shows that there is less activity than during week-days, the same holds for other week-ends.

The following section briefly describes the observations made while analysing the blackhole dataset. Table 2 (l) regroups the most occurring AS Numbers observed in the blackhole dataset. By looking up the AS Numbers, it can be observed the most occurring ASs but also the majority of all listed ASs originate from China (italic font in Table 2(l)). This can be explained for example by the high number of activities versus the proportion of all hosts in a country. On the other hand, it can be concluded that the Great Firewall[7] of China does not filter leaked packets.

N°	ASN	Frequency
1	*4134*	4 596 319
2	*4837*	1 382 960
3	*3462*	367 515
4	4766	312 984
5	*4812*	211 468
6	*9394*	166 110
7	9121	156 303
8	*4808*	153 585
9	*9318*	135 811
10	4788	116 105

N°	Most occurring "TLD[8]"s	Frequency
1	arpa	535 181
2	com	222 257
3	net	32 970
4	cn	21 920
5	org	21 567
6	cc	21 386
7	sh	17 964
8	pl	11 567
9	local	9 910
10	info	7 242

Table 2: (l) Top occurring ASNs and (r) top-occurring queried non-resolvable TLDs

Table 2 (r) shows the most occurring domain extensions in the dataset by analysing DNS A queries. Here, these DNS queries were resolved and it was checked which domains can be resolved.

N°	Anti-virus Application	Frequency
1	Kaspersky	6 005 896
2	Avast	2 906 176
3	Symantec	502477
4	McAfee	307 984
5	TrendMicro	57 004
6	Bitdefender	42 008
7	Sophos	25 023
8	AVG	22 118
9	Comodo	14 906
10	Panda	7 474

Table 3: Top occurring Anti-virus queries

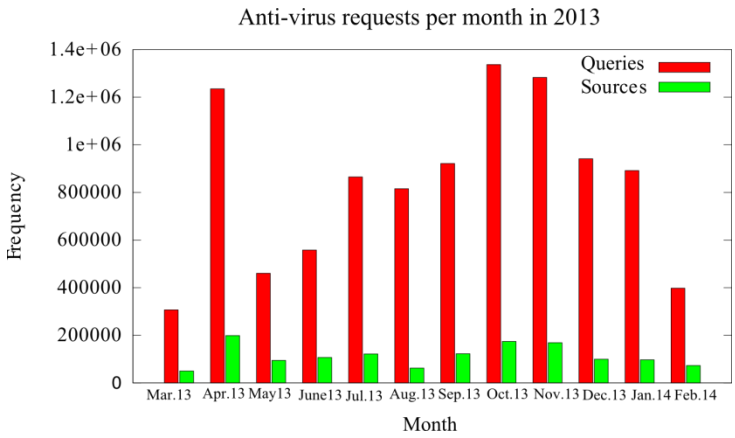

Figure 6: Anti-virus queries per month in 2013

[7] Great Firewall of China: The Golden Shield Project is a censorship/surveillance project by the Ministry of Public security in China.
[8] TLD: Top-level domain name extension.

From the domains which cannot be resolved, the domain name extension denoted by "TLD" is extracted and listed in Table 2 (r). The most looked-up domains have the .arpa domain name extension that is used for technical infrastructure purpose and used for reverse DNS look-ups. A possible explanation for this fact can be that a lot of non-resolvable domain names had a wrongly set up structure, were misspelled or even not existing.

Table 3 regroups the DNS queries for the most occurring Anti-virus applications which try to resolve their host domain names. Figure 6 shows the impact of badly configured resolvers on Anti-virus solutions deployed on hosts. Anti-virus software usually communicates with their mother companies to fetch new malware signatures or to transmit some client data for their intelligence activities/programs. The domains for queries related to anti-virus software solutions are grouped together in a set, as shown in Table 3. The sum of the queries per month per domain is represented in Figure 6. The magnitude of sources and queries is 7 times lower on average. An investigation on this confirms that the sources querying domains related to anti-virus software are quite verbose.

The 'Kaspersky' product is the most occurring product in Table 3, but a more detailed analysis shows that for the domain 'mcafee.com', 32 638 sub-domains can be identified. The majority of these queries request the 'avts.mcafee.com' subdomain and have the following structure: 'xxx.avts.mcafee.com'.
These queries are lookups from the McAfee-software to check if a suspicious file is a malicious file by generating a 32-byte fingerprint that is sent to a McAfee server[9] for their intelligence program. A similar behaviour can be observed for other anti-virus software like Avast and Sophos.

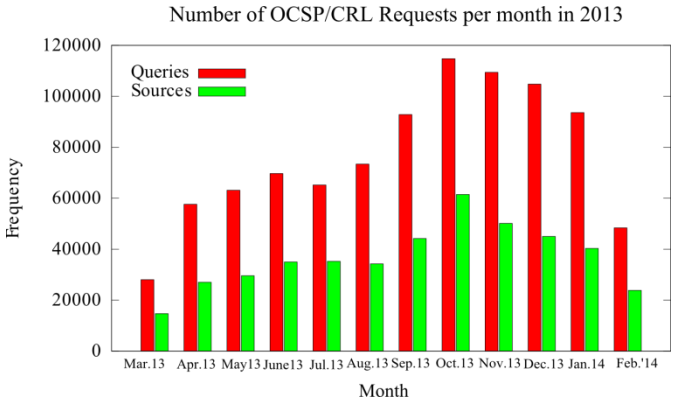

Figure 7: OCSP/CRL requests per month

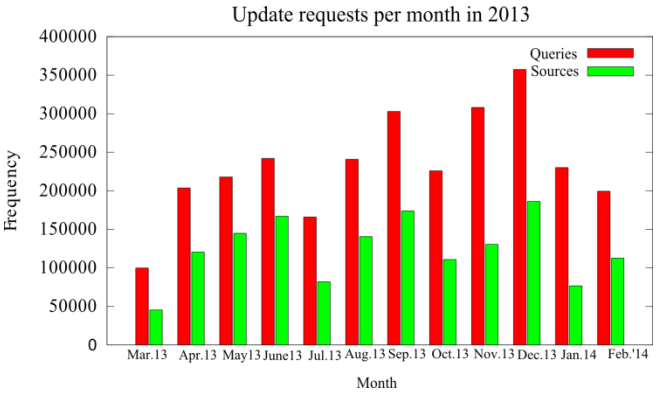

Figure 8: Software update requests

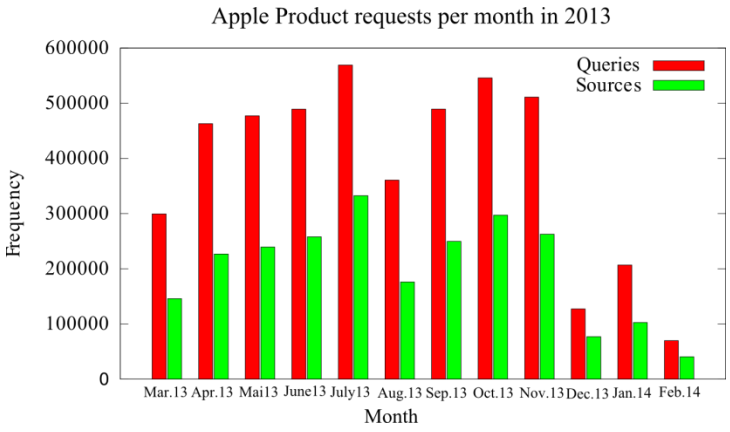

Figure 9: Apple Product requests per month

Figures 6, 7, 8 and 9 show that a badly configured resolver may have an impact on the whole security of a system. Each graph represents a group of software that might be impacted due to a bad resolver configuration. The time unit is represented in months on the x-axis. The fact of having less data on the period of Mar.13 and Feb.14 can be explained by the fact that the dataset only started the 2013-03-12 and ended on 2014-02-18.

[9] https://kc.mcafee.com/corporate/index?page=content&id=KB53735

For the columns entitled 'Queries', the frequency on the y-axis represents the total number of DNS queries that were observed in a particular month regarding a set of queries. When a host performs the same query multiple times, then these queries are also counted several times. Therefore, the columns for the frequency of 'Sources' are also given for each figure. A 'source' can be defined as a couple *src = (source IP-address; source port)*. The total number of queries for a specific query is only counted once per source to estimate the magnitude of badly configured hosts. The more, the source port is used as an identifier for the source instead of solely using the source IP, since Network Address Translation (NAT) [RFC2663] is quite common in IPv4 configurations, which means that a single public IP-address can represent more hosts. The 'Sources' columns are usually lower than the 'Queries' columns. This means that hosts usually execute more than one query.

In each figure it can be observed that the problem of badly configured resolvers is persistent over time. If more resolvers are hosted on a machine, the error may stay undiscovered for a long time as long as the second resolver works properly. However, if only one resolver is configured, the DNS queries silently hit the passive blackhole. The concerned hosts do not receive an answer and the replying service is likely not functioning properly. In both cases an operator of a malicious blackhole could infect malicious replies to control the hosts.

Figure 7 presents the amount of queries for the revocation of X.509 certificates. A commonly used protocol to fetch the Certificate Revocation List (CRL) is via the Online Certificate Status Protocol (OCSP). A host usually consults an URL to get the information about certificates, if they are revoked or obsolete for example. A badly configured DNS setting can make this technique inefficient because the service might be not available anymore. Software may assume a certificate as valid, when the CRL is not available. This may be used as an advantage for an attacker. In addition to this, an attacker may modify these lists by pretending that the revoked certificates he/she is using are still valid by removing them from the revocation list.

Figure 7 gives the number of 'Sources' and 'Queries' which are increasing, compared to the beginning of 2013. An increase in CRL activity can be observed between Sept.13 and Dec.13 by around 5%. A possible explanation is that during that period a migration period for more security in certificates was launched, such that 1 024-bit keys[10] were no longer supported after Dec 31, 2013 and a lot of certificates updated. The more, during that period a Microsoft advisory bulletin[11] suggested its users to update certificates due to a security flaw. The dataset stopped in Feb.14, therefore the number of sources and queries decreases from Jan.14 to Feb.14.

The queries given in Figure 8 show the update requests for popular software ranging from operating system updates to browser updates. Security software updates and dedicated software updates were also observed. The set holds 1 144 queries for software update domains. The sources and the number of peers differ on average by a factor of 1.95 because most programs hardly ask for updates more than twice a day. On average there are 124 223 sources that do bogus DNS requests for installing updates. A malicious blackhole operator could for instance ship malicious programs by pretending to be the update server.

An example about the usage of Apple Inc. products is given in Figure 9. On average 200 671 bad apple products were observed. However, the details of Apple Inc. queries are quite interesting. In total 5 053 queries were identified around Apple Inc. products. For the domain 'apple.com' 2 391 sub-domains were counted. The sub-domain 'xxxx.phobos.apple.com' was counted 4 161 times and refers to apple iTunes[12]. From an attacker perspective the query 'xxx-buy.itunes.apple.com' can be misused for performing Man-in-the-middle attacks to steal money from iTunes users.

The remaining queries mostly refer to content delivery networks (CDN) such as Akamai. However, a few queries do not belong to these categories and the related queries look like phishing sites for iTunes. A deeper analysis show that someone with a defective DNS resolver resolved these domains for different reasons, either a victim is 'phished' or an attacker tries to test the phishing by a defective DNS configuration.

[10] http://www.symantec.com/page.jsp?id=1024-bit-migration-faq
[11] http://technet.microsoft.com/en-us/security/advisory/2862973
[12] https://discussions.apple.com/message/20342782

4.2 **Logging information**

Besides the insights provides by DNS data, also logging information reached the blackhole sensor. Logging information from configuration files of hardware or software includes sensitive information except if encryption is used. From this dataset some observations are presented by analysing extracted logging information.

4.2.1 Leaking sensitive information

An example is the disclosing of entries of a logging system from a router, as shown in Figure 10. Similar entries are generated by printers, firewalls and other devices. A mistyping in the configuration file of a device can have undesired effects and leak internal information. In Figure 10, the example of a router configuration attempt shows, that a lot of sensitive information can be disclosed. This includes for example the type/model of the device, the IP-address, the username and even the used password. Besides the fact, that this information is publicly available, it can also be misused by a possible attacker as a resource for a more targeted attack. More information respectively other examples of leaked SYLOG or SNMP information can be seen in [Wagner13].

```
Aug 13 10:11:51 M6000-G5 command-log: [10:11:51 08-13-2012
VtyNo: vty1 UserName: XXX IP: XXX ReturnCode: 1
CMDLine: show subscriber interface gei-0/2/1/12.60
Aug 13 10:46:05 M6000-G5 command-log: [10:46:05 08-13-2012
VtyNo: vty2 UserName: XXX IP: XXX ReturnCode: 1
CMDLine: conf t]
Aug 13 10:46:10 M6000-G5 command-log: [10:46:10 08-13-2012
VtyNo: vty2 UserName: XXX IP: XXX ReturnCode: 1 CMD
Line: aaa-authentication-template 1100]
```

Figure 10: Logs from a router

4.2.2 Network reconnaissance and social engineering

Other information of the blackhole can provide relevant information for network reconnaissance. Network reconnaissance is a major step in an attack life-cycle and can be considered as a pre-attack step, since it is the task of collecting information about a potential target. Information from a blackhole can be used for example by extracting information such as NetBios machine types, machine names or wrongly configured devices disclosing information. Figure 11 only gives a subset of information extracted from the dataset.

Machine Names		NetBios machine types	
ASTTF.NET	HELP.163.COM	Frequency	Types
ASUEGYI.INFO	HP CLIENT1	1 322	Workstation
ASUS1025C	MACBOOKAIR-CAD7	105	Server
DEFAULT	DELICIOUS.COM	26	Unknown
MAIL.AFT20.COM	DELL1400	23	Browser Server
S3.QHIMG.COM	SMTP.163.COM	21	Domain Controller
Many others			Many others

Figure 11: Other information extracted from blackhole

This additional information on machine types, product types or leaked logging messages provide precious information about an entity and constitute the base for a targeted attack, either for traditional hacking or for social engineering attacks. For example, if an attacker can combine the extracted machine names with the router logs containing the username, password, product type, IP-address and domain name, he/she can initiate a social engineering attack against an entity by simply pretending on place to be hired to perform maintenance work on a specific machine since he/she knows the details about the machine and in general can search on the Internet for additional contact information about that specific entity.

4.3 **Spelling mistakes in IP-addresses**

The IP space of the monitored darkspace network part is very close to legitimate private network IP-address space. Since the monitored darkspace is a set of /24 address space, it can be assumed that there are no omissions on the first 7 positions since the IP network address has the format xyz.xyz.[0-255].[0-255], if there were omissions in these two first blocks, the corresponding traffic would not be monitored by this blackhole.

Another assumption is that there is no special sign error on a 'dot' position, such as a '.' replaced by a ',', ':', ';', etc. In general, insertion, transposition and substitution errors can occur on all positions in an IP-address, but in this evaluation, the main focus is to evaluate why traffic arrives in the blackhole.

Position	Insertion	Omission	Transposition	Substitution
1	Yes	**No**	Yes	Yes
2	Yes	**No**	Yes	Yes
3	***Yes***	***No***	***Yes***	***Yes***
4	.	**.**	.	.
5	Yes	**No**	Yes	Yes
6	Yes	**No**	Yes	Yes
7	Yes	**No**	Yes	Yes
8	.	.	.	.
9	Yes	Yes	Yes	Yes
10	Yes	Yes	Yes	Yes
12	Yes	Yes	Yes	Yes
12	.	.	.	.
13	Yes	Yes	Yes	Yes
14	Yes	Yes	Yes	Yes
15	Yes	Yes	Yes	Yes

Table 4: Possible mistyping positions in used darkspace

[Kuk92] observes that 23% of all errors occur on the 3rd character in a written word. The used network IP-address is very close to a legitimate one, where only the third position of the IP-address is changed. Since a lot of traffic arrives in the blackhole and by analysing this traffic more deeply on IP-addresses, it can be concluded that a large part of this traffic is legitimate, but that a substitution error happened on the third position of the IP-address.

It can be observed that the occurred error is a simple keystroke error, where an adjacent key was hit, as for example, instead of typing '1*23*.123.123.123'; '1*22*.123.123.123' was typed. Another possible error is a transposition error on this third word position since by transposing two adjacent keys this also deviates traffic into the blackhole, as for example '1*23*.123.123.123' into '1*32*.123.123.123'. Errors on other positions in the IP-address cannot be analysed as such, since errors may occur on all positions, even if the probability of having multiple errors in a word is very low [Kuk92], but not impossible. Table 4 regroups all possible positions where a single error may be placed in a word. Since this is work in progress, unfortunately, the prediction model for analysing the probability for more error-prone network-spaces cannot be presented here.

4.4 **Security Recommendations**

The observations summarized in this paper show that the phenomenon of misconfigured DNS and other configuration errors is omnipresent and the roots can be diversified. In this section, some security recommendations for avoiding data to end in the blackhole will be given.

In case of the data leakage due to spelling mistakes it can be said that a double checking of configuration files and a double check of used IP-addresses is recommended. Since the blackhole is close to RFC1918 networks it can be said that a lot of data is leaked into the blackhole due to these spelling mistakes due to unawareness of the configurator or by using mistyped default values. Leaked logging information as presented in Section 4.2 can be mostly avoided by restricting outgoing traffic and by adjusting settings on the Firewall. Besides this, it is recommended to test configurations first on spelling mistakes, and additionally to test them with a low log-information level to see if traffic really reaches its destination.

In section 4.2, sensitive information is leaked by a misconfigured router. To mitigate data leaks, it is recommended to use encryption to avoid clear-text passwords/sensitive data, since this data may be misused for targeted (social-engineering) attacks. The same applies to SYSLOG and SNMP information, also here, it is recommended to refer to recent versions which support encryption.

In case of the DNS, if only one name-server is configured in the DNS, the following assumptions may be made. Security systems such as anti-virus software do not properly resolve to communicate with their producers to get new updates, operating systems such as Microsoft Windows and Mac OS do not resolve properly to get updates and X.509 certificates are not properly checked. Even if a second name-server is configured correctly, the malicious potential of these configuration errors should not be underestimated. A typographic error may leak DNS traffic to a name-server under the control of an attacker, providing him/her the possibilities to send back malicious software via fake updates.
However, old filtering practises[13] could prevent these defective behaviours. The outgoing connections to the Internet should be filtered. For a given network perimeter a list of known name-servers should only be allowed for DNS resolution via UDP and TCP on port 53. All other DNS traffic should be logged. The records in these log files should be regularly audited and usually provide a good indication of badly configured or compromised machines. If filtering is not possible, an owner of a machine should review and test the DNS resolution of his or her machine by observing outgoing DNS queries. The fact of simply filtering outgoing DNS requests and allowing just a limited set of name-servers does not solve the problem where private DNS queries leak. Therefore a set of local name-servers should be set up and properly configured that does the proper distinction of the private DNS queries and the public ones. A typographic error in the configuration of a DNS resolver will not generate traffic to the Internet anymore.

In case of X.509 certificates a working certificate process is essential. A user should ensure that his/her browser supports certificate revocation and verify that this process is working. This can be done by auditing the software doing the revocation.

5. Related work

A lot of research has already been done in the area of darkspace analysis. Popular techniques used to evaluate the content of blackholes are for example packet classification [Crovella06] or time series analysis [Wustrow10], where ports, TCP flags or operating system information can be extracted. A lot of work has also been done in the area of feature distribution [Wustrow10, Zseby12 and Zseby13] where results provide insights about used attack tools for example. Other techniques for evaluating blackhole content are based on attack detection mechanisms for malware, scanning by referring to statistics or even visualisation.

Since the availability of IPv4-address space has become a scarce resource, the organisation of darkspace has become a difficult task. Therefore a possibility is to divide darkspaces into different categories [Bailey06], such as distributed darkspaces where different small address-spaces are combined, or greyspaces [Harrop05] that are sparse darkspaces, or dynamic darkspaces that represent darkspaces of temporary use. A newer trend is to apply darkspace monitoring onto IPv6 address space. In [Czyz13], an analysis is performed on unsolicited traffic in IPv6 darkspaces where results are then compared to traffic from IPv4 darkspaces.

6. Planned future work and Conclusion

Data collected by the darkspace sensor highlights a lot of precious information and shows that spelling mistakes in networking can have fatal consequences while leaking highly sensitive information. Other errors can be attributed to simple copy-errors from configuration manuals. These observations raise the question, if some networks are more probable to errors, such as spelling errors, mistyping or editing errors as others. A work in progress is the extension of the present evaluation method to all kind of IP-address. A new prediction model for spelling errors in networking tasks is implemented in order to detect a spelling mistake or to identify an abuse. This prediction model uses the probabilities of possible errors on given positions in a word and takes as an additional factor the probabilities of occurrences of digits in a number, but this is still work under submission. Another future work is to apply the method of predicting errors in darkspaces on IPv6 address space and to check the probability of spelling mistakes for IPv6-addresses and to draw possible conclusions between spelling mistakes in IPv4 and IPv6-addresses for example.

[13] http://books.google.lu/books/about/Firewalls_and_Internet_Security.html?id=XI52je-zaW8C&redir_esc=y

This analysis of a darkspace shows that simple spelling mistakes or other misconfiguration can leak private data and constitute a major threat to network security. A lot of the data captured by the probe can be used as attack vectors for a hacker to plan malicious activities against a network and its security. Security recommendations are given for administrators and users such as verification of configuration files and monitoring of critical services such as DNS, SYSLOG and other configuration files. Since recent past, the monitoring of such IP darkspaces has become a major source of information in cybersecurity and has been become a tool in early IT-warning system.

Acknowledgments

We would like to acknowledge the whole RESTENA, RESTENA-CSIRT and CIRCL team for their support for realizing this work.

References

Ahmed09. Ahmed A., Andrew C. and Mohay G. 2009. Characterising Anomalous Events Using Change-Point Correlation on Unsolicited Network Traffic. Identity and Privacy in the Internet Age, (5838), pp.104-119. Springer.

Bailey05-1. Bailey M., Cooke E., Farnam J., Provos N., Posaen K. and Watson D. 2005. Data reduction for the scalable automated analysis of distributed darknet traffic. Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement, pp. 21-21.

Bailey05-2. Bailey M., Cooke E., Jahanian F., Nazario J. and Watson D., 2005. The Internet Motion Sensor - A Distributed Blackhole Monitoring System. In Proceedings of Network and Distributed System Security Symposium (NDSS '05), pp. 167-179.

Bailey06. Bailey M., Cooke E., Jahanian F., Myrick A. And Sinha S. 2006. Practical Darknet Measurement. Proceedings of the Annual Conference on Information Sciences and Systems. Pp. 1496-1501.

Crovella06. Crovella M., Balachander K. 2006. Internet Measurement. Infrastructure, Traffic and Application. Wiley Press.

Czyz13. Czyz J., Lady K., Miller S.G., Bailey M., Kallitsis M. and Karir M. 2013. Understanding IPv6 Internet Background Radiation. Proceedings of the 2013 Conference on Internet Measurement Conference, ACM, pp. 105-118.

Netbeacon. Dulaunoy A. Netbeacon Monitoring Your Network Capture. Available through: <https://github.com/adulau/netbeacon/> [Accessed 4 March 2014]

PassiveDNS. Dulaunoy A. and Tricaud S., Scrutinizing a Country using Passive DNS and Picviz or how to Analyze big dataset without losing your mind. Available through: <https://github.com/adulau/pdns-toolkit/blob/master/slides/picviz-pdns.pdf > [Accessed 4 March 2014].

Gates37. Gates A.I. 1937. A list of spelling difficulties in 3786 words: Showing the "hard" spots. Bureau of publications Teachers College, Columbia University, 166 pages.

Harrop05. Harrop W. and Armitage G. 2005. Defining and Evaluating Greynets. Proceedings of IEEE Conference on Local Computer Networks. pp. 344-350.

Kuk92. Kukich K. 1992. Techniques for Automatically Correcting Words in Text. ACM Computer Surveys, 24 (4), pp. 377-439.

Poll83. Pollock J.J., Zamora A. 1983. Collection and characterization of spelling errors in scientific and scholarly text. Journal of the American Society for Information Science, 34 (1), pp. 51-58.

RFC1918. Rekhter Y., Moskowitz B., Karrenberg D., de Groot G.J. and Lear E., 1996. RFC 1918: Address allocation for private Internets, Available through: <https://tools.ietf.org/html/rfc1918> [Accessed 4 March 2014].

RFC2663. Srisuresh P. and Holdrege M. IP Network Address Translator (NAT) Terminology and Considerations. Available through: < http://tools.ietf.org/html/rfc2663> [Accessed 12 April 2014]

Vinod04. Vinod Y., Barford P. and Plonka D. 2004. On the design and use of Internet sinks for network abuse monitoring. In Proceedings of Recent Advances in Intrusion Detection. Springer Verlag, pp. 146-165.

Wagner13. Wagner C., Dulaunoy A., Wagener G. and Stiefer M. 2013. Another Perspective to IP-Darkspace Analysis. Available through: <http://www.circl.lu/files/tf-csirt-first2013-circl-restena-blackhole.pdf> [Accessed 4 March 2014]

ASMATRA. Wagner C., François J., State R., Dulaunoy A., Engel T. and Massen G. 2013. ASMATRA: Ranking ASs providing transit service to malware hosters. In Proceedings of the Integrated Network Management Symposium (IM 2013), pp. 260-268.

Wustrow10. Wustrow E., Karir M., Bailey M., Jahanian F. and Huston G. 2010. Internet Background Radiation Revisited. In Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC '10), ACM, pp. 62-74.

Zalewski05. Zalewski M., 2005. Silence on the Wire, a Field Guide to Passive Reconnaissance and Indirect Attacks. No Starch Press.

Zseby12. Zseby T. and Claffy K. 2012. Workshop report: darkspace and unsolicited traffic analysis (DUST 2012). SIGCOMM Computer Communication Review 42, ACM, pp. 49-53.

Zseby13. Zseby T. 2013. IP Darkspace Analysis. Advances in IT early warning, Frauenhofer IRB Verlag, pp. 21-29.

Biographies

Alexandre Dulaunoy encountered his first computer in the eighties, and he disassembled it to know how the thing works. While pursuing his logical path towards information security and free software, he worked as senior security network consultant at different places (e.g. Ubizen, now Cybertrust). He co-founded a startup called Conostix specialized in information security management, and the past 6 years, he was the manager of global information security at SES, a leading international satellite operator. He is now working at the national Luxembourgian Computer Security Incident Response Team (CSIRT) in the research and operational fields. He is also lecturer in information security at Paul-Verlaine University in Metz and the University of Luxembourg. Alexandre enjoys working on projects where there is a blend of "free information", innovation and a direct social improvement. When not gardening binary streams, he likes facing the reality of ecosystems while gardening or doing nature photography.

Gérard Wagener has a bi-national Ph.D. in computer science at the University of Luxembourg and INPL Nancy, France. He was working for 4 years in the global information security team at SES, a leading international satellite operator. He is working for the national Computer Security Incident Response Team (CSIRT) coordination center for Luxembourg. His doctoral research focuses on adaptive decoying systems to improve intelligence gathering on attackers in computer networks. He is the founder and lead developer of the Adaptive Honeypot Framework, which serves as the solid foundation for constructing intelligent honeypots. Gérard comes from the malware research community, where he worked on projects such as sandboxes for monitoring and analyzing malicious software. In addition to these hands-down activities, his scientific work has investigated malware classification using phylogenetic trees and intelligent high-interaction honeypots driven by game theory. In his spare time Gérard can often be found in his garage fixing cars and doing metalwork.

Marc Stiefer holds an Engineer Degree from the former University of Luxembourg in computer science. He actually works as system and security engineer for the RESTENA Foundation and is a member of the Computer Security Incident Response Team (CSIRT) of RESTENA. In his spare time, Marc enjoys running marathon.

Cynthia Wagner holds a Ph.D. in computer science from the University of Luxembourg. She actually works as a research engineer for the Computer Security Incident Response Team (CSIRT) of the RESTENA Foundation, the NREN of Luxembourg. Besides this, she works in the security team for DNS-LU, the registry for the .lu ccTLD. In her spare time, she is a senior lecturer in computer and network security at the University of Luxembourg and likes travelling.