# Assignment Block 2 - Spamhaus

Hugo Bijmans [4253760]
Jorrit van der Spek [4174348]
Ben Hup [1150065]
Eveline Pothoven [4380509]
Lisette Altena [1526413]
Group 8

October 3, 2016

## 1 Introduction

This document is a report for the WM0824TU Economics of Cyber Security course at Delft University of Technology. The purpose of this report is to give the reader an overview of the cyber security metrics with the focus on botnets and their related SPAM. The report addresses various security issues, gives an overview of the ideal (but not always possible) metrics for security decision makers and summarizes the existing metrics defined in other studies. The second part of this report is about putting those metrics into practice by using them to investigate a given data set. In this case, this is the Spamhaus CBL data set, which lists blocked IP-addresses due to botnet infected spamming. An overview of desired metrics is given, followed by the evaluation of them.

## 2 Security issues

People are becoming increasingly dependent on information and information systems. On the one hand, these systems create a lot of potential, but a major downside is the amount of risk it creates. Each system or piece of information could fail, leak, get lost or be stolen. Cyber criminals, who are the ones trying to steal the information and create disruption, often use botnets to do harm.

### 2.1 What is a botnet?

A bot is a malicious piece of software, which is – without the knowledge of the computer's user – installed at someone's machine. The bot enables the communication with the command and control (C&C) infrastructure. A botnet is a collection of compromised machines running the same bot program. The network consists of between thousand and one million machines, which all receive instructions and commands from the bot's master. The master, who administrates the C&C server, can use these machines for several criminal activities such as spamming (Elliot, 2010).

## 2.2 Methods

Since the costs of running the zombies, like electricity and bandwidth, are covered by the victims, the costs for the criminals are very low. Meanwhile the reward can be very high. These rewards can either be money, espionage or political protest opportunities (Elliot, 2010). Therefore criminals use different methods, which will be briefly described in this section.

### Distributed Denial of Service Attacks

A DDoS attack refers to an assault on the availability of a system emanating from multiple sources. It can result in degradation of the service being attacked, for example slower response times, or complete unavailability of the service. The motive of the attacker can range from a political protest to targeted attacks on a business or individual for criminal extortion attempts (Elliot, 2010).

### Spam

Email is considered spam if it is unsolicited and sent in bulk. Besides junk mail from businesses to advertise goods, emails containing viruses are also considered spam. Another category of spam mail sends the receiver to websites that contain scripts to collect information for the purpose of identity theft and other criminal operations. The mail could also contain links that claim to take you of the mailing list, but in fact the intention is to verify whether the email is actively used.

### Click fraud and adware

Botmasters receive money from adware companies for downloading adware applications on a zombie in the botnet, up to $0.20 for each download. Furthermore, botmasters generate revenue by installing software which automatically clicks on internet advertisements. They receive a compensation for each time the program clicks on the advertisement, so-called pay per click schemes.

### Phishing scams

Phishing is a form of internet fraud with which the criminal tries to acquire sensitive information, by masquerading as a trustworthy person or business. The victim has to click on a link to a fake, but real-looking, website that requires to fill in personal information. The criminal can use this information or will sell the information.

### Ransomware

Criminals are able to encrypt the victim's hard drive. The Readme.txt which is left on the computer, tells the victim how to contact them to purchase the decryption key. Since many home users rarely back up their hard drives, they will probably pay the amount of money.

### 2.3 Impact

The impact for the victims varies. If a government's or company's system appears not to be as impenetrable as they should be, it can cause reputational damage. Secondly the attacks can cause disruption of varying degrees. For example a productivity loss, because of deleting spam or dealing with a localized malware infection. The time and money is taken away from other priorities (McDermot, 2006). Especially for the public, electronic crime may result in emotional damage. For example the loss of data, like irreplaceable photos, if malware has rendered a computer unusable. Lastly the botnets have a financial impact. First of all caused by the productivity loss and the time it takes to recover. Furthermore processing time, electricity and bandwidth is used and people have to invest in anti-virus and firewall software.

## 3 Ideal metrics for security decision makers

In this chapter the ideal metrics for security decision makers will be discussed. First there will be explained what these metrics are. Then the key security decision makers regarding botnets will be mentioned with their ideal metrics. Lastly there will be a conclusion about these metrics.

Measuring the security level is not a straight-forward task. To get good knowledge about the level of security a lot of factors should be considered, these factors are called metrics. Every security investment model builds on security metrics which define the model's inputs, outputs, and parameters (Böhme, 2010). Metrics can be categorized into four types; metrics based on controls, vulnerability, incidents and prevented losses. Metrics based on controls and vulnerability are cost driven and therefore deterministic, instead of the metrics based on incidents and prevented losses. They are driven by events caused by attackers and thus have a stochastic nature. A good model for accurate measuring should be built on metrics that contain all four types but in practice security decision models consist mostly on metrics based on control and vulnerability. This is because they are closest to cost and easy to measure. On the other hand metrics based on incidents and prevented losses are hard to measure and require a lot of resources. In this section we will not focus on the practicability and costs of these metrics but solely on what would be ideal metrics for the security decision makers.

There are a lot of decision makers dealing with security issues regarding botnets. They all have a different point of view on these security issues and thus have different ideal metrics. These five key decision makers will be discussed below; the users, the internet service providers, the criminals, the governments and the security industry.

First of all there are the users who receive the spam sent by botnets. The users can be separated into private and corporate users. For both the metric; the amount of spam they receive compared to the average amount of spam, can help them understand if they are targeted. The metrics; income loss by spam, productivity loss by spam, average cost of security breaches and the amount of infected machines in their network are of great value to decision makers in companies.

The internet service providers would have the following ideal metrics to help

the make the best security decisions. Which IP-addresses from their clients sent spam, which IP-addresses from their clients receive spam and which botnet are they part of? This will tell them which users they have to block. They can compare these metrics with other Internet services providers to indicate if their security measures are adequate.

A metric that can be used by the criminals, who use botnets to send spam are; success rate of spam bombs. With this metric criminals can pinpoint weaknesses in security and exploit it. Metrics that tell them if their command and control nodes have been adjusted to change their botnet into a sinkhole would also be of great use to the criminals. The amount of competing botnets with their relative size can be used for economic purposes.

Governments are interested in metrics that give them better understanding in the activities of the cyber criminals and the economic losses they cause. They can use the following metrics to accomplish this; the amount of infected machines active in their country, which botnets are controlled in their country, the specific botnets that are targeting their country and the economic damage caused by botnets.

Finally, there is the security industry. Their main purpose is to sell security solutions to customers. The security industry would use metrics that enables them to take down botnets and filter spam effectively. Ideal metrics would be; IP-addresses and location from infected machines, locations of command and control nodes, what botnets are currently active, the amount of infected machines and the content of the spam.

In conclusion the ideal metrics for security decision makers consist of metrics that would help them to eliminate all risk. However this is an impossible prospect because the ideal metrics are either extremely difficult to implement or there are no incentives to invest in these metrics.

# 4 Existing metrics

Security metrics are very important. Nowadays, the economic climate does not allow spilling resources for information security: they are limited. The security spending must be justified and allocated. Therefore, the right metrics are necessary. If one invests a lot in information security, he wants to get actual security, and reap certain benefits (Gañán, 2016). But the concept 'information security' is broad. Defining the metrics for this project - about bots operating on computers, sending malicious malware - could help to specify this concept.

If one takes a look at current literature, more and more articles are written about metrics and information security: an upcoming field. Already in 2008, Zhuang was talking about metrics in the world of spam. His metrics mainly focused on botnets, and listed three metrics (Zhuang et al., 2008):

1. Capability of botnet controllers: estimate the total size of each botnet based on their 9 days of observation in the experiment
2. Level of activity (botnet): estimate the active working set of each botnet in a short time window, such as one hour. Think of the spam sent (such as the number of spam emails) per botnet
3. Active size (botnet): the number of machines/IPs used for sending spam email messages by this botnet during this short time window

If one looks a bit further in literature, you might find a surprising amount of information. But there is one very useful paper, summarizing all this information and insights about spam metrics. Moura and Van Eeten (2015) listed a summary of current botnet metrics. First, they pointed out what the requirements of useful metrics are, such as 'comparative over time' and 'comparability'.

Secondly, they carried a literature review on the current metrics, and proposed a classification of these metrics into three categories:

1. IP-based: metrics using the originating IP address of traffic related to infected machines
2. Host-based: metrics based on data that directly and reliably indentifies individual hosts on the internet.
3. Proxy-based: metrics that are estimations based on traffic volume associated with botnets (Moura & van Eeten, 2015)

These categories are shown in the second column of the table. The other columns quite speak for themselves, the last three might need some more explanations. The categories presented above can be further extended: by aggregation (per country for example), by normalization, or by ranking: being turned in a rating, based on a different scale than the original metric.

Table 1: summary of current botnet metrics (Moura & van Eeten, 2015)

| Metric | Type | Meas. Window | Data source | Agg. | Normalized | Ranking |
|---|---|---|---|---|---|---|
| estimated # of hosts | IP | per hour/per day | Sinkhole | - | - | |
| extrapol. # of bots | IP | per day | Honeynet and Darknet | # of Source ASes | Avg, number of IP scanned per botnet | |
| # of bots per AS | IP | per day | Spam email | ASes, BGP | | Top 20 AS and countries sending spam |
| Malscore | IP | 60 days | IRC-based botnets HTTP-based botnets | ASes | Size of AS | AS Ranking |
| Botnet activity | IP | per day | Spam data | ISP | # of subscribers per ISP | ISPs |
| CCM | Host | quarter | Malwares cleaned | Country | # of computers cleaned/ 1000 unq computers executing the MSRT | Countries |
| Unique malicious objects | Host | quarter | Malwares detected | Country, % of unique attacked users | | Countries |
| Spam volume | Host | quarter | spam, web exploits, malware, DDoS | Themes for spam, platform (Windows, Linus, Mobile) Country | | Countries, Platform |
| # bot IDs per countries | Host | 10 days, per hour, per day | Sinkhole | Country | | Countries |

| Metric | Type | Meas. Window | Data source | Agg. | Normalized | Ranking |
|---|---|---|---|---|---|---|
| Suspicnss. score | Proxy | per day | recursive DNS (RDNS), spam | | | |
| # of malicious domains | Proxy | 1.2 days | DSN spam | | | |
| Active size | Proxy | per day | Spam emails | Clustered emails into spam campaigns/# of countries participated in sending spam | | |
| Badness score | Proxy | per day | Click-spam | Search Ad Network, Mobile Ad Network, Contextual and Social Ad Networks | | |
| ASrank | IP | per day | malware | ASes | Size of AS | AS |
| Max spam vol./asn Min spam vol/asn | IP | per day | spam | ASes, country | Size of AS | Country |
| % malicious hosts per asn | IP | 30 days | Phishing, malware, spam | ASes | Size of AS | % of malicious hosts per asn |

| Metric | Type | Meas. Window | Data source | Agg. | Normalized | Ranking |
|---|---|---|---|---|---|---|
| % spam caught | IP | per day | spam | ASes | Size of AS, Size of subnet | reputation, subnet reputation asn |
| cluster based reputation | IP | per day | spam emails | BGP prefix cluster, DSN cluster | | |
| % spam caught | IP | per day | spam | ASes | Size of AS, Size of subnet | reputation, subnet reputation, asn |
| # of infected domain clusters | Proxy | per day | DNS | DNS cluster | | |
| # of bots per timezone | IP | per day | sinkhole | bots per continent | total number of bots | # of syn connections by botnet sent per continent |
| # of unique IP per spam campaign | IP | per hour | spam emails | Countries, ISPs | | Top 20 countries with the most Bot IPs, Top 20 IPSs that host the most bot IPs |
| # of unique suspected bots | IP | per day | sinkhole | flows (src IP, dest IP, src IP, src port, dest port) | | |

## Reflection of current metrics

To reflect on the current metrics might be the most important part of reviewing spam and botnet metrics. What are the issues with those current metrics, or do they work perfectly fine? The shortcomings of the metrics will be discussed in the same categories as used before.

IP-based metrics violate several requirements (such as reliability) due to DHCP and NAT effects. For example, it is possible three bots are operating, from three different laptops, behind a single public router IP address. This shows

it is very complex to count botnet presence in ISP network: the IP addresses do not correspond to the number of operating bots.

Host-based metrics are known as more reliable than IP metrics and proxy metrics. The data used for these metrics is very precise, but this is exactly the problem. The data requires access to the hosts themselves, but the access to this data is either restricted or presented to the public in aggregated levels. These metrics are very reliable, but it is hard to obtain the necessary data.

Proxy based metrics are not very precise. This occurs because they mainly express estimates on the number of infected machine, they do not express actual data. It would not be a big problem, if the estimation could me made precisely. Unfortunately, there are many factors influencing the measurements, which make the estimation unreliable. Proxy based metrics are not completely useless, but should be used with caution, and only for purposes that fit with their shortcomings (Moura & van Eeten, 2015).

# 5 Metrics from dataset

In addition to the data in the given data set additional metrics can be defined for insight in the spamming behavior. The metrics defined in this chapter can be derived solely from the data in the data set. When a metric can only be derived when additional external data is available this will be stated explicitly. The derived metrics are as follows:

## 5.1 Unique IP addresses controlled by botnet

Each botnet consists of a Command and Control node that controls all the bots in a botnet. The number of unique IP addresses under control of the Command and Control node signifies the efficiency at which a botnet can achieve the goal of e.g. spamming. It is to be noted that multiple devices can be connected behind a single IP address via Network Address Translation (NAT). The limitation of accuracy due to NAT is noted by for practical reasons ignored in this paper. Metric interesting for: Internet Service Providers (ISP) to know what botnets are most prevalent to know where support is necessary most for cleaning customers. Also Internet security companies can use the metric to know what software developments add most value to products.

## 5.2 Top 10 country per botnet

Botnet signatures can be counted per country. Bots are not bound by geographical of national boundaries. Within each country operate a certain amount of bots that belong to a certain botnet which can be counted. For the top 5 biggest botnets a top 10 of countries is established to show what botnets are best represented in which country. The top 5 biggest botnets metric is established from the first metric "Unique IP addresses controlled by botnet". Metric interesting for: ISP's can learn whether they overall have many infected customers and should have more controls in effect to counter customers getting infected. Internet security companies write software for the world, focusing on one country would be a less profitable endeavor. Governments can learn from his metric if advising the public on safe Internet use is effective relative to other countries.

## 5.3 Top 10 Internet Service Providers (ISP) that host botnets

Devices are connected to the Internet via an Internet Service provider (ISP), including bots and botnet Command & Control nodes. Certain Internet Service providers (ISP) could be more likely to host bots than others. This metric allows for the insight which ISPs host the most bots in a top 10 list. Metric interesting for: ISP's can learn whether they have many infected customers overall and should have more controls in effect to counter customers getting infected. Internet security companies can profit from this metric by offering their services to ISP's. The government could use this metric to give incentive to ISP's to clean up and support customers is keeping their devices clean.

## 5.4 Top 10 SPAM sending countries

Not all countries send the same amount of SPAM, nor considering SPAM sent per capita. There is no homogeneous set of laws that is international applicable. Neither are digital criminality laws enforced with the same magnitude, which results in disparities in SPAM sent per country. A top 10 of sent SPAM per country gives a metric that can be used to decide which countries pose an additional security risk. Metric interesting for: ISP's could enforce stricter email policies for countries via a high SPAM rate. Security companies can use the metric to know what countries need expertise the most and offer expertise. Governments know how their country performs in protecting against SPAM relative to other countries.

## 5.5 Botnet activity per country

Each country experiences a different amount of active bots that send SPAM. For each country separately the top 10 of most active botnets can be calculated. This metric is valuable to consider for a company in a certain country whether to invest in countermeasures against the top botnets. Metric interesting for: ISP's can merely learn an aggregated view of botnet activity. Internet Security companies can learn to what countries they can offer their expertise besides the top 10 in metric 5. Governments not showing in the top 10 of metric 5 can learn from this metric how well they relatively perform.

## 5.6 Number of countries active for the top 10 botnets

For the top 10 biggest botnets in IP count the number of different countries can be calculated. This metric allows for insight in how dispersed botnets are over different countries. Especially when SPAM emails contain links to infect more devices this metric could give insight in what countries a certain botnet is not effective in gaining more bots. Metric interesting for: ISP's can learn what botnet poses the biggest threat to customers. Internet security companies can learn on what botnet to focus software development of focus gaining knowledge. Governments know on what botnets to create public awareness and pose the biggest threat.

## 5.7 SPAM activity by time frame

SPAM activity is not the same for a 24 hour period. The world population is awake at different times around the world. This metric gives insight in the amount of SPAM sent per hour in a 24 hour cycle (i.e. one earth day). If there is a difference in SPAM sent per hour this could mean that SPAM sending processes are not fully automated but require human intervention. Metric interesting for: ISP's can calculate for which time frame maybe additional personnel is necessary for support, as the SPAM activity and thus infection rise. The same point can be made for Internet security companies and governments.

## 5.8 SPAM sent via Tor node

SPAM is in most cases sent directly from a bot's IP address. However, it might be possible that certain bots use the Tor network to send SPAM. This metric

gives insight in how much percent uses the Tor network to send SPAM to remain truly anonymous. Metric interesting for: This metric is less interesting for ISP's at their own as they can't do anything against using Tor. However, Internet security companies and governments could use this metric to get insight in how SPAM is delivered to victims. The government could enact policies against Tor in collaboration with Internet Security companies and ISP's.
There is one more advanced metrics to be considered, but requires additional data sets to be generated. The eleventh metric could be geographically locating the Command & Control nodes of the botnets. By using IP address to GPS coordinate conversion the geographical locations and time stamps of sent SPAM could be used in combination with considering a spike in sent SPAM. During a spike different bots must have gotten a command from the botnet Command & Control (C&C) node to send SPAM. The latency between bots and botnet C&C could be used to roughly estimate where the botnet C&C is located in the world. Because this metric requires coupling multiple data sets together and could have a large inaccuracy due to rough estimation of latency and ignoring that packets could be routed in sub-optimal paths this metric is not pursued in this paper.
This chapter showed that at least ten metrics could be retrieved from the Spamhaus SPAM data set. It is explained what value can be retrieved from the metrics and what and how actors could retrieve value from the metrics. Both quantitative numeric metrics can be produced as well as metrics, such as a heat map, that appeal to getting a quick grasp of SPAM production.

# 6 Evaluation of defined metrics

## 6.1 Methodology

The given data set was built of 10.554.552 rows and 8 columns. Before analyzing the data set, some cleaning had to be done. Firstly a random row in the middle of the data set, containing the names of the columns, was removed. Next, all the records which did not contain a time stamp or an ASN number were removed. At the end 15.636 records were removed (which is 0.14 % of all the data in the data set). The final data set contains 10.538.915 to work with. SPSS was used to clean the data, R was used to analyze the data.

## 6.2 Metrics

**Amount of nodes added per botnet**

In the Spamhaus Database, every IP address is unique, because the database is used by providers to block those addresses that send spam. It's interesting to see which botnets made a major contribution to the block list in the sample period.

As seen in the table and pie chart, one botnet (the BOT C_conficker) is far more present in the data set than all the others. The Conficker botnet was first detected in November 2008 and became the biggest botnet of all times, infecting more than 9 million computers at its peak in 2009 (Neild, 2009). Nowadays, the botnet is abandoned by its makers, but the computers are still infected, which makes them vulnerable for attacks which use the Conficker infrastructure to gain access to the computer.

Table 2: Contribution to the block list per botnet

| Rank | Botnet | number of IP addresses | % of total |
|---|---|---|---|
| 1 | BOT c_conficker | 3.654.641 | 35% |
| 2 | MPD | 920.926 | 9% |
| 3 | BOT dyre | 818.051 | 8% |
| 4 | BOT gamut | 705.992 | 7% |
| 5 | BOGUS | 610.905 | 6% |
| 6 | BOT c_confickerab | 404.182 | 4% |
| 7 | BOT c_zeroaccess | 369.457 | 4% |
| 8 | BOT s_tinba | 301.139 | 3% |
| 9 | BOT s_zeus | 258.807 | 2% |
| 10 | BSIP | 242.669 | 2% |

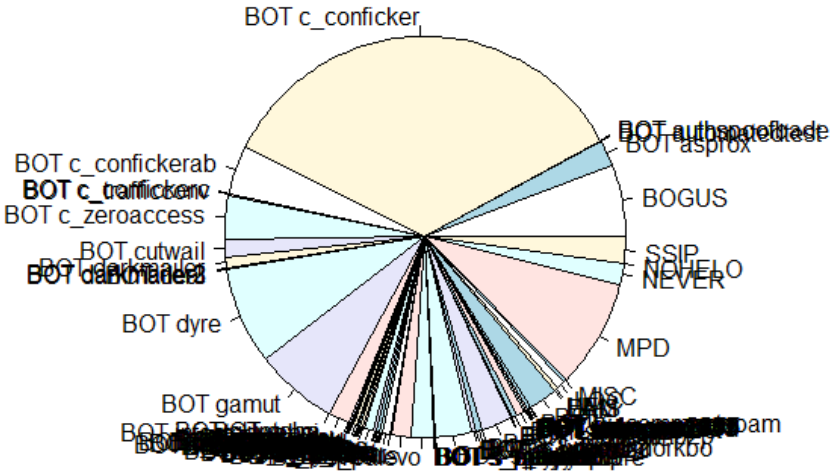


Figure 1: Pie chart of the contribution of different botnets

### Top 10 providers hosting Botnets

Investigating the ASN codes present in the data set, we were able to see which providers hosted the most infected computers who were sending spam. The ASN codes are resolved using the RIPE Database (RIPE, 2016), which translate the ASN codes to provider names. It's not a surprise to see that China's backbone provider is the host of the most IP addresses, since this is also one of the countries which host the most spam sending IP addresses.

The pie chart is showing a wide variety of providers involved in hosting malicious IP addresses. The big organizations, like the Chinese, Vietnamese and Indian backbone are obvious the most present, but the major part of the blocked IP addresses belong to other providers. So, besides China, Vietnam and India, there is not really a provider hosting a large of amount of malicious IP addresses.

| # | ASN Number | Name | Country | Number of records |
|---|---|---|---|---|
| 1 | AS4134 | CHINANET-BACKBONE | China | 701.205 |
| 2 | AS45899 | VNPT-AS-VN | Vietnam | 693.424 |
| 3 | AS9829 | BSNL-NIB | India | 501.901 |
| 4 | AS17974 | TELKOMNET-AS2-AP | Indonesia | 286.744 |
| 5 | AS7552 | VIETEL-AS-AP | Vietnam | 200.668 |
| 6 | AS45595 | PKTELECOM-AS-PK | Pakistan | 198.701 |
| 7 | AS18403 | FPT-AS-AP | Vietnam | 177.056 |
| 8 | AS4837 | CHINA169-Backbone | China | 167.542 |
| 9 | AS3462 | HINET | Taiwan | 143.430 |
| 10 | AS8151 | Uninet S.A. de C.V | Mexico | 133.791 |

Table 3: Top 10 providers hosting botnet nodes

**Top 10 countries of blocked IP addresses**

Each IP address in the data set is mapped to a specific country. Investigating the country field in our data set gives the following results, visible in table 4.

Normalizing this data was hard, due to the lack of information about the number of computers present in a country. In order to achieve some sort of normalization, the number of active internet users per country was taken. This number does not accurately represent the number of personal computers, since this also includes mobiles phones and tablets with an internet connection. But with that taken into account, it gave more insight to the data than without normalization.

Table 4 shows the effect of normalization. China for example is ranked second in having the most records of blocked IP's in this data set, but the normalized value ranks it on the 9th position in the list, thanks to the massive amount of computers present in this Eastern country. Notable countries are Pakistan, Iran and of course Vietnam. These countries host relatively the most malicious IP addresses, when the number of internet users is used to normalize the values.

Table 4: Top 10 countries of blocked IP addresses

| # | Country code | Country | Number of records | % of total | no. of internet users (Wikipedia, 2016) | normalized | normalized # |
|---|---|---|---|---|---|---|---|
| 1 | VN | Vietnam | 1.162.444 | 11% | 40,597,779 | 2,863319198 | 1 |
| 2 | IN | India | 1.152.149 | 10% | 462,124,989 | 0,249315451 | 8 |
| 3 | CN | China | 1.098.155 | 10% | 721,434,547 | 0,152218244 | 9 |
| 4 | RU | Russia | 579.619 | 5% | 102,258,256 | 0,566818781 | 5 |
| 5 | BR | Brazil | 480.780 | 5% | 120,111,118 | 0,400279348 | 7 |
| 6 | ID | Indonesia | 357.036 | 3% | 80,000,000 | 0,446295 | 6 |
| 7 | IR | Iran | 285.827 | 3% | 25,074,125 | 1,139928113 | 2 |
| 8 | US | United-States | 268.296 | 3% | 286,942,362 | 0,093501705 | 10 |
| 9 | IT | Italy | 252.711 | 2% | 35,942,120 | 0,703105437 | 4 |
| 10 | PK | Pakistan | 250.058 | 2% | 34,342,400 | 0,728131988 | 3 |

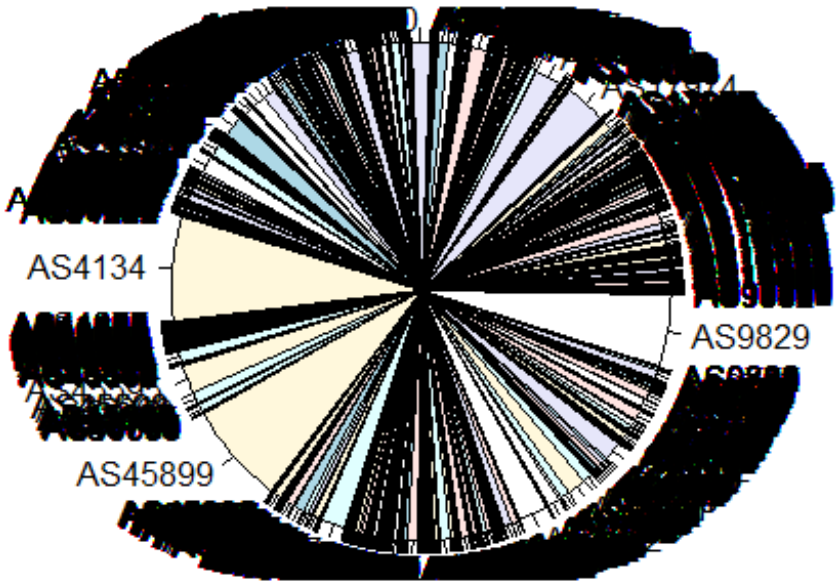


Figure 2: Distribution of providers hosting botnets

**Top 10 countries per botnet**

Does every botnet target the same country? Or will some botnet focus on the one country and another on another country? This question is answered in this section using a simple top 10 ranking of countries. In order to achieve this overview, the country and botnet information is aggregated and a massive table is constructed. This table is further analyzed in Excel and used to produce the list as seen in table 5.

The country preferences of the biggest 10 botnets in the data set were analyzed and they all had the focus on the same countries, which are visible in table 5. Notable is the fact that a country like Peru is present in the list, which wasn't in the other comparisons. The rest of the list is not a surprise, with

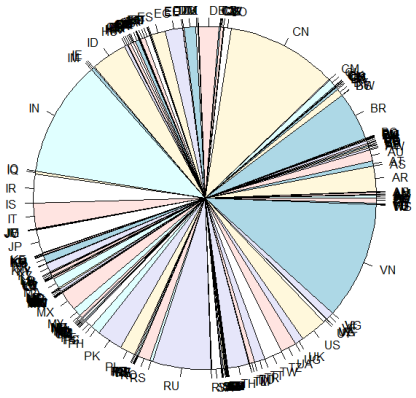


Figure 3: Distribution of providers hosting botnets

India, China and Pakistan as the major contributors.

When the information in table 4 is combined with this information, something really interesting is clearly visible. Almost every botnet infected PC in India is infected by one of the 10 biggest botnets in the data set. This in contrast to Vietnam, where the major players account for only 22% of the malicious IP's. This means that the biggest botnets are using bots located in India, but more smaller botnets use Vietnamese infected computers to perform spam attacks.

Table 5: Top 10 countries of the 10 major botnets

| # | Country | Number of records |
|---|---|---|
| 1 | India | 1.260.240 |
| 2 | China | 908.768 |
| 3 | Pakistan | 630.764 |
| 4 | Russia | 412.524 |
| 5 | Iran | 344.396 |
| 6 | USA | 310.828 |
| 7 | Vietnam | 265.432 |
| 8 | Mexico | 246.115 |
| 9 | Peru | 210.043 |
| 10 | Brazil | 198.788 |

Table 6: Top 10 countries per botnet

| # | c_conflicker | MPD | dyre | gamut | BOGUS |
|---|---|---|---|---|---|
| 1 | India | India | India | India | India |
| 2 | China | China | China | China | China |
| 3 | Pakistan | Pakistan | Pakistan | Pakistan | Pakistan |
| 4 | Russia | Russia | Russia | Russia | Russia |
| 5 | Iran | Iran | Iran | Iran | Iran |
| 6 | USA | USA | USA | USA | USA |
| 7 | Vietnam | Vietnam | Vietnam | Vietnam | Vietnam |
| 8 | Mexico | Mexico | Mexico | Mexico | Mexico |
| 9 | Peru | Peru | Peru | Peru | Peru |
| 10 | Brazil | Brazil | Brazil | Brazil | Brazil |

| # | c_conflickerab | c_zeroaccess | s_tinba | s_zeus | BSIP |
|---|---|---|---|---|---|
| 1 | India | India | India | India | India |
| 2 | China | China | China | China | China |
| 3 | Pakistan | Pakistan | Pakistan | Pakistan | Pakistan |
| 4 | Russia | Russia | Russia | Russia | Russia |
| 5 | Iran | Iran | Iran | Iran | Iran |
| 6 | USA | USA | USA | USA | USA |
| 7 | Vietnam | Vietnam | Vietnam | Vietnam | Vietnam |
| 8 | Mexico | Mexico | Mexico | Mexico | Mexico |
| 9 | Peru | Peru | Peru | Peru | Peru |
| 10 | Brazil | Brazil | Brazil | Brazil | Brazil |

**Botnets active per country**

Different botnets are active in different countries. As the following tables show, the conficker botnet is present in every country, but some botnets are more present in one country than in every other. The p2pzeus bot is significantly more active in Italy than in the rest of the world.

Table 7: Botnets active per country

| # | Vietnam | India | China | Russia | Brazil |
|---|---|---|---|---|---|
| 1 | c_conflicker | gamut | c_conflicker | c_conflicker | c_conflicker |
| 2 | BOGUS | c_conflicker | dyre | MPD | MPD |
| 3 | dyre | MPD | MPD | dyre | BOGUS |
| 4 | MPD | dyre | s_tinba | BOGUS | dyre |
| 5 | c_zeroaccess | s_zeus | BOGUS | c_conflicker | c_zeroaccess |
| 6 | c_conflicker | BOGUS | SSIP | c_zeroaccess | c_conflicker |
| 7 | kelihos | BSIP | gamut | asprox | s_zeus |
| 8 | asprox | s_tinda | c_conflicker | NEVER | asprox |
| 9 | s_zeus | SSIP | NEVER | kelihos | gamut |
| 10 | gamut | asprox | c_zeroaccess | s_tinda | kelihos |

| # | Indonesia | Iran | United States | Italy | Pakistan |
|---|---|---|---|---|---|
| 1 | c_conflicker | c_conflicker | c_conflicker | c_conflicker | c_conflicker |
| 2 | dyre | MPD | MPD | dyre | MPD |
| 3 | MPD | dyre | BOGUS | MPD | dyre |
| 4 | BOGUS | BOGUS | dyre | c_zeroaccess | s_tinba |
| 5 | c_zeroaccess | c_conflickerab | c_conflickerab | c_conflickerab | BOGUS |
| 6 | c_conflickerab | c_zeroaccess | c_zeroaccess | BOGUS | c_conflickerab |
| 7 | s_zeus | asprox | kelihos | gamut | asprox |
| 8 | gamut | BSIP | asprox | s_p2pzeus | NEVER |
| 9 | kelihos | kelihos | gamut | asprox | c_zeroaccess |
| 10 | aprox | NEVER | s_tinba | kelihos | cutwail |

## 6.3 SPAM activity by time frame

SPAM activity is not the same for a 24 hour period. The world population is awake at different times around the world. What kind of influence would that have on the blocking of certain IP-addresses by Spamhaus?

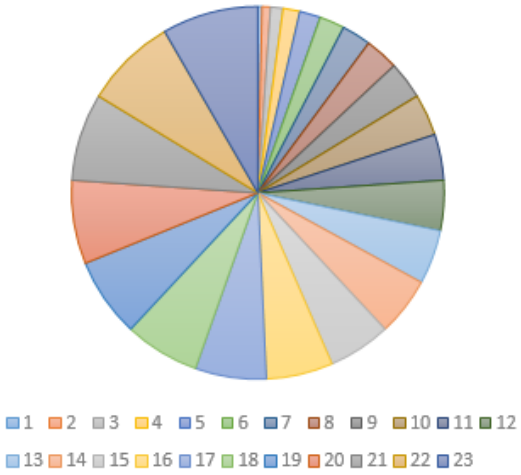


Figure 4: Distribution IP's blocked by time frame

This pie chart shows the answer to the question stated above very quickly. The distribution of blocking malicious IP-addresses is more a continuous process than a discrete one. However, in the early hours of the night and morning (from 01:00 till 12:00h) just a quarter of the block occur. It slowly builds up to 23:00h when the most blocks take place.

# 7 Conclusions

The purpose of this report is to analyze the Spamhaus data set, and learning the importance of measuring cyber security and the challenges to create meaningful metrics. From the analysis of the Spamhaus data set significant anomalies have been found that can be used for more informed decision making and improve cyber security processes such as preventive controls. Analyzing providers hosting the greatest number of botnet infected PC's results in the fact that CHINANET-BACKBONE (AS4134) hosts over 19% botnet infected PC's (701.205 out of 3.654.641). Furthermore, the analysis shows that the c_conflicker botnet has globally the largest installed base representing 35% of all infected PC's. Also, India, China and Pakistan harbor the biggest the largest amount of infected bots for the top botnets such as c_conflicker, MPD, dyre, gamut, BOGUS and others. However, India is the only country where almost every botnet infected PC harbors a botnet infection from the top 10 botnets as established in this report. Also, by observing the total number of infected PC's India overshadows all other countries in absolute numbers or almost a third of the total data set. Not all of the defined metrics are evaluated. Due to the large data set, we were unable to analyze the botnet activity via Tor nodes. This is an interesting topic for further research. The results of this report can contribute to more effective security policies of ISP's, Internet security companies and governments knowing what botnets have the largest installed base and what countries pose an elevated probability of being a source of attack.

# References


[1] Elliot C.(2010)., *Botnets: To what extent are they a threat to information security?*. Information Security Technical Report, 15(3), 79-103.

[2] McDermot J. (2006, February 8)., *SPAM - The Issues, Impact and Reducing SPAM (Part 1)*. Retrieved from `http://www.windowsecurity.com/whitepapers/anti_spam/Impact-Reducing-SPAM-Part1.html`.

[3] Böhme, Rainer (2010)., *Security Metrics and Security Investment Models, Advances in Information and Computer Security, Lecture Notes in Computer Science Volume 6434, pp 11.* Retrieved from `https://www.is.uni-muenster.de/security/publications/Boehme2010_SecurityInvestment-IWSEC.pdf`.

[4] Gañán C. (2016) *WHat to measure* Retrieved from `https://edge.edx.org/courses/course-v1:DelftX+WM0824+Fall_2015/courseware/655640221aaf4d54b4502a86b0746514/6a99f22f340347d89469f89232114c4d/`.

[5] Zhuang L., Dunagan J., Simon D., Wang H., Tygar J. (2008)., *Characterizing botnets from email spam records*. USENIX Association, 1-9.

[6] Moura G, van Eeten M. (2015)., *Documentation of botnet metrics methodology and development*. Advanced Cyber Defence Centre, 1-95.

[7] Neild, Barry. (2009-01-16), *Downadup Worm exposes millions of PCs to hijack*, CNN, retrieved from `http://edition.cnn.com/2009/TECH/ptech/01/16/virus.downadup/?iref=mpstoryview`.

[8] RIPE NCC. (2016). *RIPE Database Query*. Visited 30-09-2016 at `https://apps.db.ripe.net/search/query.html#resultsAnchor`

[9] Wikipedia.org (2016). *List of countries by number of Internet users*. Visited 30-09-2016 at `https://en.wikipedia.org/wiki/List_of_countries_by_number_of_Internet_users`

# A Appendix

## A.1 R commands used

To analyze the given Rstudio was used. In this appendix all the R code is presented.

```
# Set working path and import data
setwd("~/TU Delft/Collegejaar 2016-2017/Economics of Cybersecurity")
spamdata <-read.csv("SpamHause_SPSS_Final.csv", header = TRUE, sep=";")

# Code to generate the distribution of the number of IP addresses per botnet
summary(spamdata$Diagnostic)

# Code to generate the piechart of botnet distribution
pietable <- table(spamdata$Diagnostic)
lbls <- paste(names(pietable))
pie(pietable, labels = lbls)

# Code to generate the distribution of providers
summary(spamdata$ASN)

# Code to generate the piechart of provider distribution
pietable2 <- table(spamdata$ASN)
lbls <- paste(names(pietable2))
pie(pietable2, labels = lbls)

# Code to generate the distribution of providers
summary(spamdata$Country)

# Code to generate the piechart of provider distribution
pietable3 <- table(spamdata$Country)
lbls <- paste(names(pietable3))
pie(pietable3, labels = lbls)

# Code to generate the data for countries per botnet and countries per botnet
aggregate(Country ~ Diagnostic, summary, data=spamdata)

# Code to process the timeframe data in Perl
use strict;
use warnings;
use diagnostics;

my %h;

print "hoi";
#open(FH, "<d:\\tmp\\cbl2.csv")  || die;
open(FH, "<D:\\Education\\\TU Delft 2016-2017 MoT
plus Leiden Univ\\Q1 WM0824TU Economics of Cyber Security
\\SPAM assignment data files from Feyzullah Cetin 2016-09-16
\\cbl.diagnosis-parsed-18092015.csv")  || die;
```

```
#my @a = ();
#while(my $line=<FH>)
foreach my $line (<FH>)
{
#  $line =~ s/\n|\r//gi;
  if ( $line !~ m/^\d+$/ )
  {
    # $line =~ s/\n|\r//gi;
    #  print $line;
      my @a = split(/\,/, $line);
    #  print "a6";
    #   print $a[6]."\n";
       my ($sec, $min, $hour, $mday, $mon, $year, $wday, $yday, $isdst) =
       localtime($a[6]);
#       print $hour."\n";
#       if(!defined($h{$hour})){ $h{$hour} = 0;}
       $h{$hour}++;
  }
}

#print "2";

foreach my $key (sort keys %h){
        print "$key: ".$h{$key}."\n";
}

# Code to generate the pie chart timeframe
install.packages("data.table")
require(data.table)
setwd("D:/tmp/")
#Of zelf headers toevoegen aan csv, of lijn hieronder FALSE zetten
spamdata <- read.csv("cbl2.csv", header = TRUE, sep = ",")
#names(spamdata) <- c("IP","ASN","Allocation","Country","Domain","State",
"time_t","Diagnostic")
data <- data.table(spamdata)
summary(spamdata$Diagnostic)
data[,list(diag=Diagnostic,hour=format(as.POSIXct(time_t,
origin="1970-01-01"), format = "%H")),by="time_t" ]
pietable <- table(data$hour)
lbls <- paste(names(pietable))
pie(pietable, labels = lbls)
```