Assignment Block 2

SpamHaus

DRAFT

Group 8
Jorrit van den Spek, Hugo Bijmans, Eveline Pothoven, Ben Hup, Lisette Altena

WM0824 Economics of Security

Table of Contents

1. Security Issues

Spam, it is a major source of frustration for internet users. More than halve of the e-mail traffic consists of spam (Figure 1). It causes high costs for companies. Employees who could otherwise spend their time on more productive work, waste their time. Network errors, which are among others caused by employees responding to spam mail, largely impacts the productivity. Furthermore spam uses storage and bandwidth, which could otherwise be used for useful purposes, spam filters might unfairly block emails, because of which emails are delayed. (http://www.windowsecurity.com/whitepapers/anti_spam/Impact-Reducing-SPAM-Part1.html)

Besides the fact that spam impacts businesses by wasting a lot of time, it affects the environment as well. Spam expert Richi Jennings calculated together with climate change consultant ICF International the environmental impact of spam. According to the study, the energy consumed in transmitting and deleting spam last year (62 trillion) is similar to the amount of electricity used in 2.4 million U.S. homes. (https://resources2.secureforms.mcafee.com/LP=2968).

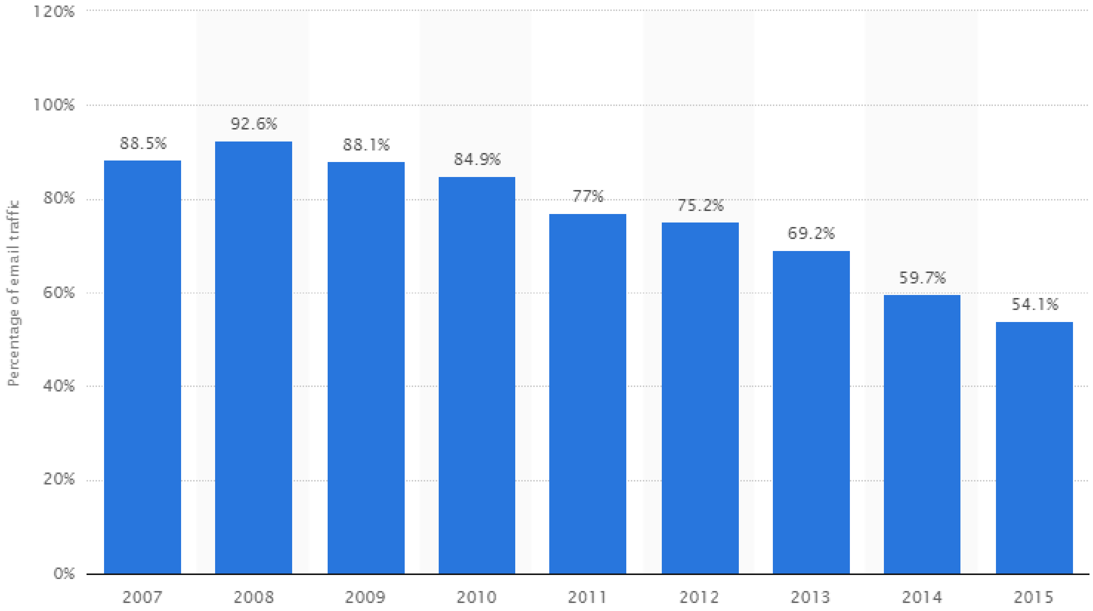


Figure 1: Global spam volume as a percentage of total e-mail traffic from 2007 to 2015, source: https://www.statista.com/statistics/420400/spam-email-traffic-share-annual/

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.

2. Ideal metrics for security decision makers

The metrics for security decision makers should consist of metrics based on all four types. In practice metrics are usually based on control and a bit on vulnerability. This is because controls are closely related to the cost and are put in place to mitigate risk. Metrics based on vulnerabilities evaluate how these controls preform under threat. These metrics are deterministic in contrast to the metrics based on incidents and the prevision of loss, whom are based on events driven by attackers and are therefore stochastic. They map the losses whenever a curtain events occurs. Implementing these metrics is resource consuming and are used less.

There are a lot of decision makers dealing with the issue of spam. In the next section the most important decision makers will be mentioned with specific metrics that are useful to them.

First of all there are the users who receive the spam. The users can be separated into private and corporate users. For both the matric: ratio of spam received to average amount of spam, can help them understand if they are targeted. The matric; income loss by employees engaging in spam and income loss by false negatives is also of great value to decision makers from companies. The internet service providers can use the following metric to help the make security decisions. The amount of spam received by their clients. They can compare this matric 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 is; success rate of spam bombs. With this metric criminals can pinpoint weaknesses in the security and exploit it. The amount of competing botnets with their relative size can be used for economic purposes. Governments are interested in metrics that calculate economic losses and the amount of damage done by engaging in spam. Last there are the email software developers, they would use metrics to indicate if their platform is targeted and has weak spots. Metrics like the amount of false positives and amount of spam send to their domain could help the software developers.

In conclusion the ideal metrics for security decision makers consist of a mix of already existing metrics from the four types and specific metrics suited for the activities different security decision makers have.

3. 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 (bron: online lecture 2.2). But what 'information security' is a wide concept. In this project, the main focus is Spam, regarding the project SpamHaus.

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, 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

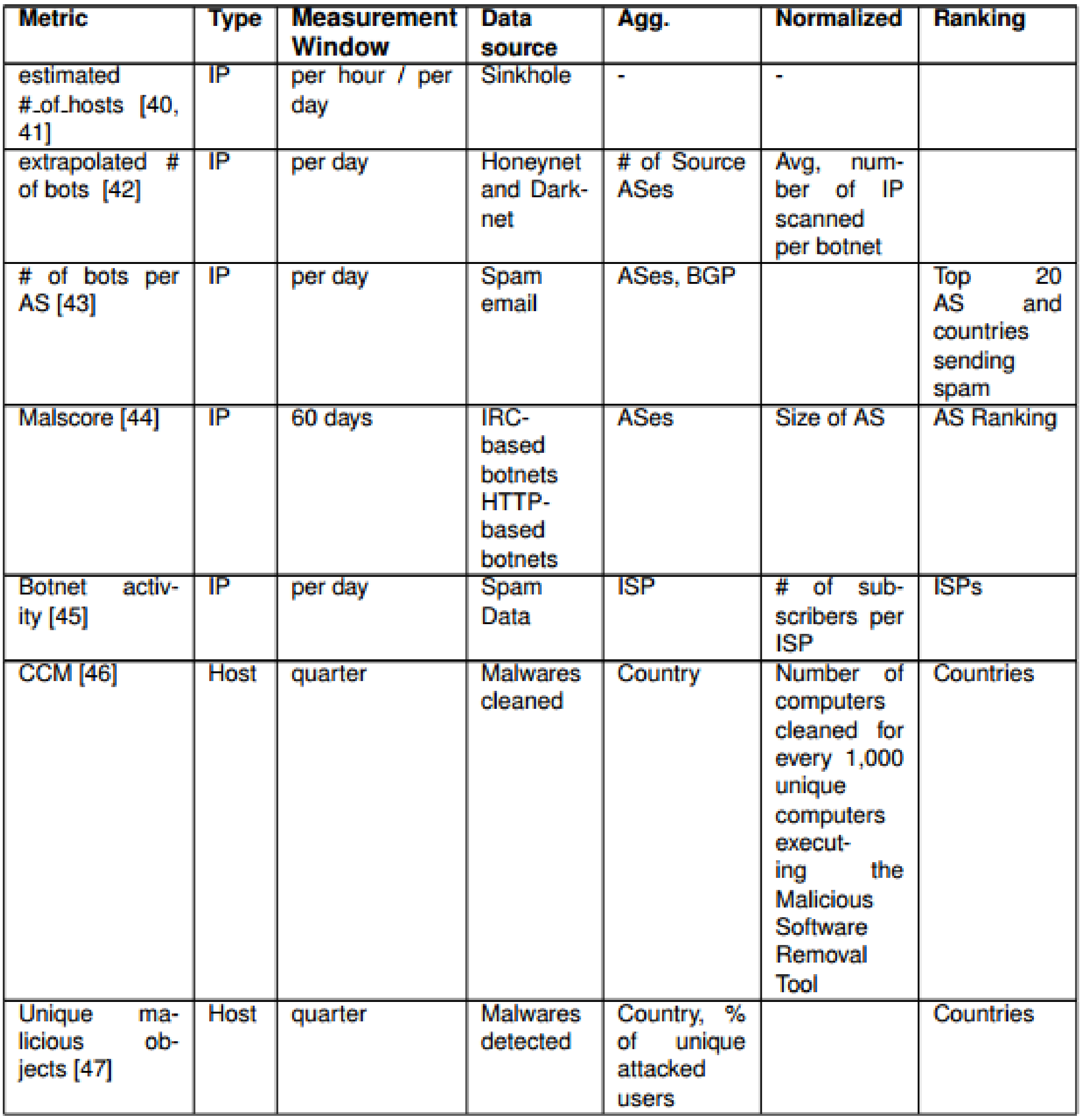
If one looks a bit further in literature, you might find a surprising amount of information. But there is one very useful paper, summarizing all this information and insights about spam metrics. Moura and Van Eeten (2015) listed a summary of current botnet metrics. First, they pointed out what the requirements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

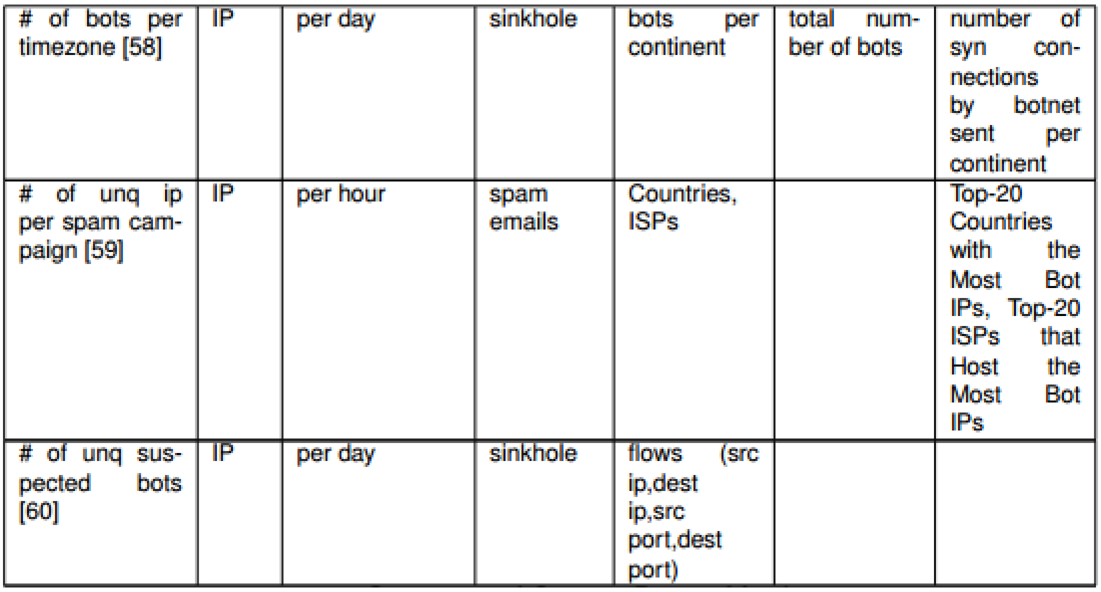
These categories are shown in the second column of the table. The other columns quite speak for themselves, the last three might need some more explanations. The categories presented above can be further extended: by aggregation (per country for example), by normalization, or by ranking: being turned 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	Measurement Window	Data source	Agg.	Normalized	Ranking
estimated #_of_hosts [40, 41]	IP	per hour / per day	Sinkhole	-	-	
extrapolated # of bots [42]	IP	per day	Honeynet and Darknet	# of Source ASes	Avg, number of IP scanned per botnet	
# of bots per AS [43]	IP	per day	Spam email	ASes, BGP		Top 20 AS and countries sending spam
Malscore [44]	IP	60 days	IRC-based botnets HTTP-based botnets	ASes	Size of AS	AS Ranking
Botnet activity [45]	IP	per day	Spam Data	ISP	# of subscribers per ISP	ISPs
CCM [46]	Host	quarter	Malwares cleaned	Country	Number of computers cleaned for every 1,000 unique computers executing the Malicious Software Removal Tool	Countries
Unique malicious objects [47]	Host	quarter	Malwares detected	Country, % of unique attacked users		Countries

spam_volume [48]	Host	quarter	spam, Web Exploits, Malware, DDoS	Themes for spam, Platform (Windows, Linux, Mobile) Country		Countries, Platform
# bot IDs per countries [12]	Host	10 days, per hour , per day	Sinkhole	Country		Countries
Suspiciousness score [49]	Proxy	per day	recursive DNS (RDNS), spam			
# of malicious domains [50]	Proxy	1.2 days	DNS , spam			
Active Size [51]	Proxy	per day	Spam emails	Clustered emails into spam campaigns / # of countries participated in sending spam		
Badness score [52]	Proxy	per day	Click-spam	Search Ad Network, Mobile Ad Network, Contextual and Social Ad Networks		
AS_{rank}[53]	IP	per day	malware	ASes	Size of AS	AS
max_spam _vol_per_asn min_spam _vol_per_asn [25]	IP	per day	spam	ASes, Country	Size of AS	Country
%_malicious _hosts _per_asn [54]	IP	30 day	Phishing, malware, spam	ASes	Size of AS	% of malicious hosts per asn
% spam caught [55]	IP	per day	spam	ASes	Size of AS, Size of subnet	reputation _subnet reputation_asn
cluster based reputation [56]	IP	per day	spam emails	BGP prefix cluster , DNS cluster		
% spam caught [55]	IP	per day	spam	ASes	Size of AS, Size of subnet	reputation _subnet reputation_asn
# of infected domain clusters [57]	Proxy	per day	DNS	DNS cluster		



# of bots per timezone [58]	IP	per day	sinkhole	bots per continent	total number of bots	number of syn connections by botnet sent per continent
# of unq ip per spam campaign [59]	IP	per hour	spam emails	Countries, ISPs		Top-20 Countries with the Most Bot IPs, Top-20 ISPs that Host the Most Bot IPs
# of unq suspected bots [60]	IP	per day	sinkhole	flows (src ip,dest ip,src port,dest port)		

Reflection of current metrics

To reflect on the current metrics might be the most important part of reviewing spam metrics. What are the issues with those current metrics, or do they work perfectly fine? The shortcomings of the spam 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).

4. Metrics from dataset

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

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.

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".

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.

4. Top 10 SPAM sending countries

   Not all countries send the same amount of SPAM, nor considering SPAM sent per capita. There is no homogenous 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.

5. Top 10 most active botnets and competition

   Botnets send a certain amount SPAM per timeframe denoting the activity of a botnet. Besides the size of a botnet sending SPAM, the activity (i.e. sent SPAM per timeframe) denotes the total amount of SPAM generated. This is another metric to show risk of a botnet.

6. 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.

7. 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.

8. Choropleth map or heat map to show clustering of SPAM sending bots

   A choropleth of heat map can show clusters of SPAM activity on a world map. For this visual metric to materialise the need for an IP-to-Coordinates database is necessary. The converted IP addresses into GPS coordinates allows for pinpointing an IP on a world map visually showing clustering of SPAM sending bots.

9. SPAM activity by timeframe

   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.

10. 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 bot's 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.

There is one more advanced metrics to be considered, but require additional datasets 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 datasets 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.

5. Evaluation of the defined metrics

Methodology

Dataset was built of 10.554.552 rows and 8 columns. Before analysing the dataset, some cleaning had to be done. Firstly a random row in the middle of the dataset, containing the names of the columns, was removed. Next, all the records which did not contain a timestamp 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 dataset contains 10.538.915 to work with. SPSS was used to clean the data, R was used to analyse the data.

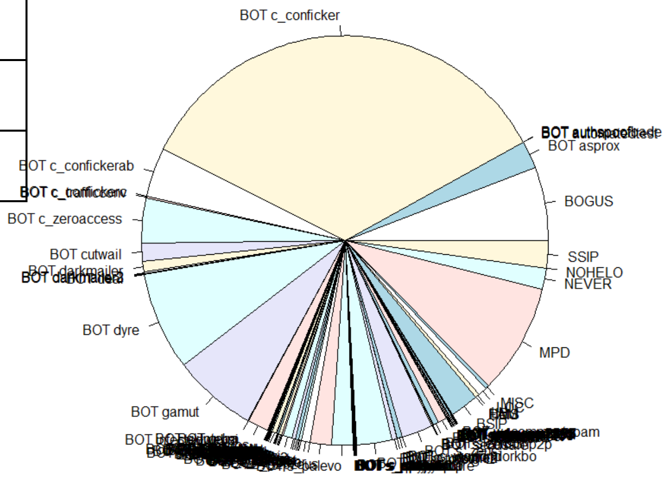
Metrics

Unique number of IP addresses per botnet

Since every IP address is unique in this dataset, this is a very straightforward question to answer. Using this R command, we were able to check which botnets contained the most IP addresses:

```
summary(spamdata$Diagnostic)
```

Rank	Botnet	#IP addresses	% of total
1	BOT c_conficker	3654641	35%
2	MPD	920926	9%
3	BOT dyre	818051	8%
4	BOT gamut	705992	7%
5	BOGUS	610905	6%
6	BOT c_confickerab	404182	4%
7	BOT c_zeroaccess	369457	4%
8	BOT s_tinba	301139	3%
9	BOT s_zeus	258807	2%
10	BSIP	242669	2%



Top 10 countries hosting Botnets

Investigating the ASN codes present in the dataset, 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 at https://apps.db.ripe.net/search/query.html#resultsAnchor

```
summary(spamdata$ASN)
```

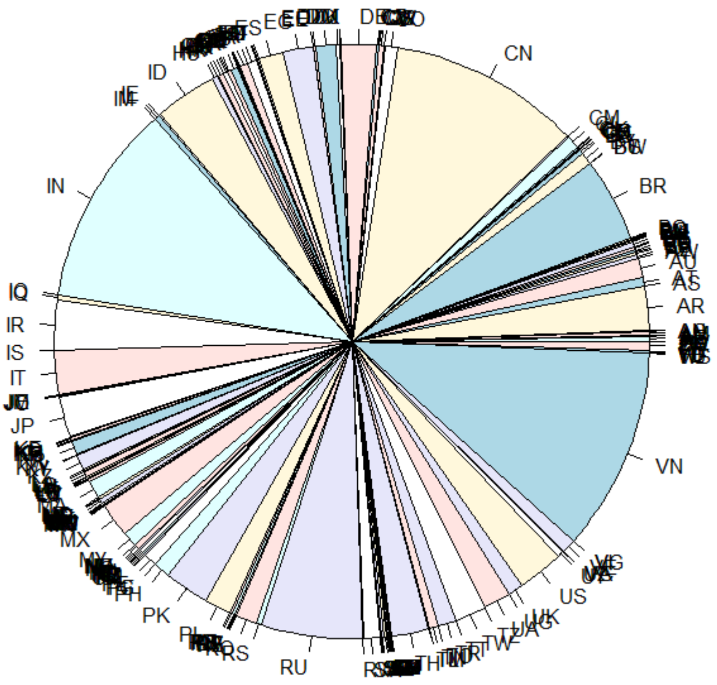
Rank	**ASN Number**	**Name**	**Country**	**#records**
1	AS4134	CHINANET-BACKBONE	CHINA	701.205
2	AS45899	VNPT-AS-VN	VIETNAM	693.424
3	AS9829	BSNL-NIB	INDIA	501.901
4	AS17974	TELKOMNET-AS2-AP	INDONESIA	286.744
5	AS7552	VIETEL-AS-AP	VIETNAM	200.668
6	AS45595	PKTELECOM-AS-PK	PAKISTAN	198.701
7	AS18403	FPT-AS-AP	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

Top 10 countries sending SPAM

Investigating the country field in our dataset gives the following results. Achieved by using this query:

```
summary(spamdata$Country)
```

Rank	Country Code	Country	#records	% of total
1	VN	Vietnam	1162444	11%
2	IN	India	1152149	10%
3	CN	China	1098155	10%
4	RU	Russia	579619	5%
5	BR	Brazil	480780	5%
6	ID	Indonesia	357036	3%
7	IR	Iran	285827	3%
8	US	United-States	268296	3%
9	IT	Italy	252711	2%
10	PK	Pakistan	250058	2%



Top 10 countries per botnet

The 10 most popular botnets are operated from the same countries. In this table, you'll see an overview. This means that the biggest botnets are using bots located in India, but more smaller botnets use Vietnamese infected computers to perform spam attacks.

```
aggregate(Country ~ Diagnostic, summary, data=spamdata)
```

	BOT c_conficker	**MPD**	**BOT dyre**	**BOT gamut**	**BOGUS**
1	Country.IN	Country.IN	Country.IN	Country.IN	Country.IN
2	Country.CN	Country.CN	Country.CN	Country.CN	Country.CN
3	Country.PK	Country.PK	Country.PK	Country.PK	Country.PK
4	Country.RU	Country.RU	Country.RU	Country.RU	Country.RU
5	Country.IR	Country.IR	Country.IR	Country.IR	Country.IR
6	Country.US	Country.US	Country.US	Country.US	Country.US
7	Country.VN	Country.VN	Country.VN	Country.VN	Country.VN
8	Country.MX	Country.MX	Country.MX	Country.MX	Country.MX
9	Country.PE	Country.PE	Country.PE	Country.PE	Country.PE
10	Country.BR	Country.BR	Country.BR	Country.BR	Country.BR

	BOT c_confickerab	**BOT c_zeroaccess**	**BOT s_tinba**	**BOT s_zeus**	**BSIP**
1	Country.IN	Country.IN	Country.IN	Country.IN	Country.IN
2	Country.CN	Country.CN	Country.CN	Country.CN	Country.CN
3	Country.PK	Country.PK	Country.PK	Country.PK	Country.PK
4	Country.RU	Country.RU	Country.RU	Country.RU	Country.RU
5	Country.IR	Country.IR	Country.IR	Country.IR	Country.IR
6	Country.US	Country.US	Country.US	Country.US	Country.US
7	Country.VN	Country.VN	Country.VN	Country.VN	Country.VN
8	Country.MX	Country.MX	Country.MX	Country.MX	Country.MX
9	Country.PE	Country.PE	Country.PE	Country.PE	Country.PE
10	Country.BR	Country.BR	Country.BR	Country.BR	Country.BR

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.

```
aggregate(Country ~ Diagnostic, summary, data=spamdata)
```

	Vietnam	**India**	**China**	**Russia**	**Brazil**
1	c_conficker	gamut	c_conficker	c_conficker	c_conficker
2	BOGUS	c_conficker	dyre	MPD	MPD
3	dyre	MPD	MPD	dyre	BOGUS
4	MPD	dyre	s_tinba	BOGUS	dyre
5	c_zeroaccess	s_zeus	BOGUS	c_confickerab	c_zeroaccess
6	c_confickerab	BOGUS	SSIP	c_zeroaccess	c_confickerab
7	kelihos	BSIP	gamut	asprox	s_zeus
8	asprox	s_tinba	c_confickerab	NEVER	asprox
9	s_zeus	SSIP	NEVER	kelihos	gamut
10	gamut	asprox	c_zeroaccess	s_tinba	kelihos

	Indonesia	**Iran**	**United States**	**Italy**	**Pakistan**
1	c_conficker	c_conficker	c_conficker	c_conficker	c_conficker
2	dyre	MPD	MPD	dyre	MPD
3	MPD	dyre	BOGUS	MPD	dyre
4	BOGUS	BOGUS	dyre	c_zeroaccess	s_tinba
5	c_zeroaccess	c_confickerab	c_confickerab	c_confickerab	BOGUS
6	c_confickerab	c_zeroaccess	c_zeroaccess	BOGUS	c_confickerab
7	s_zeus	asprox	kelihos	gamut	asprox
8	gamut	BSIP	asprox	s_p2pzeus	NEVER
9	kelihos	kelihos	gamut	asprox	c_zeroaccess
10	asprox	NEVER	s_tinba	kelihos	cutwail

6. References

Not yet here, will be in the final version!