Research Project 1: In Depth Abuse Statistics

Student Project
Michiel Timmers & Arthur van Kleef
michiel.timmers@os3.nl - arthur.vankleef@os3.nl

System and Network Engineering, Class of 2008-2009
Universiteit van Amsterdam, Netherlands

July 3, 2009

1 Abstract

This paper describes how abuse statistics are presented and what can be done to improve these statistics. Furthermore we have looked at how network characteristics are reflected in abuse load, and if this will allow us to weigh abuse statistics per network in correlation to their size, setup and use.

2 Acknowledgments

We would like to thank Carel van Straten from Spamhaus[1] and Jan-Philip Velders from Universiteit van Amsterdam[2] and SURFcert[3] for their assistance, feedback and tips during this research project. Their knowledge was very useful.

Furthermore we want to thank the Composite Blocking List (CBL)[4] for making there dataset publicly available on which we did most of our data analysis. And thanks to Team Cymru for providing a rsync to there IP to ASN service [5] in order for us to perform large amounts of lookups.

Contents

3 Introduction

3.1 General description of the project

Through the use of several different blocking lists (e.g. Composite Blocking Lists (CBL)[4] and Not Just Another Bogus List (NJABL)[6] which publish IP addresses of known spam source hosts, malicious Internet hosts can be identified.

We will monitor these lists and keep track of listed IP addresses (time listed, no. of times added/removed from list), the resulting data will allow us to identify 'problem' IP subnets. Once these subnets have been identified we will investigate the characteristics (size, setup and use) of these networks. This will give us insight in the relationship between network characteristics and abuse, and allow us to weigh, qualify and quantify the 'abuse load' of a specific network.

The term abuse describes a broad spectrum of abusive operations (port scanning, hacking, spamming, etc.) on the Internet, in our research we use the term abuse to describe the act of sending unsolicited e-mail messages (spam).

3.2 Goal and research questions

We started this project to gain more insight in the networks that are reported for sending unsolicited e-mail messages. In order for us to reach this goal we defined the following research questions:

- **Is there a relationship between the network and its abuse?** Is it possible to correlate network characteristics like size, setup, and use to the amount of abuse that takes place in a network? And will this correlation allow us to weigh the abuse load per network?
- **Influence of keeping a history** DNS blacklists like the Composite Blocking List (CBL) are publicly available, but only contain current data. We will keep a history of the additions and removals to the CBL and try to investigate if this will reveal new insights about abused networks.
- **Abuse per network** Publicly available data on abusive networks (e.g. Top 200 Spammers by Country[7] keep track of most abusive networks per country, or most give lists of countries spam originated from. We want to consider network size (number of hosts) too in these statistics. This will hopefully also lead to a way to score a network's abuse load.
- **Do IP masquerading techniques compromise CBL integrity?** We are curious to see if techniques that hide the a host's actual IP address impose a threat to CBL integrity. Especially the Network Address Translation (NAT) technique draws our attention, since this is a widely deployed method to preserve public IPv4 addresses.

3.3 Outline of this report

In section 4 of this report we will look at current statistics and how they can give a wrong impression. Section 5 will describe how we setup our research and will show some basic statistics (i.e. top n countries and networks). In subsection 5.3 broadband penetration is compared to abuse statistics, and in 5.4 we have looked at dutch ISP's and their abuse listings and handling.

In section 6 we propose grading methods to identify abusive networks. We conclude this paper with our findings and what we think can be done in further research.

4 Current statistics

The problem with abuse statistics is that many of them only show the number of abuse issues that are originated within a country or network, no further statistics are given about size, setup or use of that particular country or network. The column of table 1 lists the number of spam messages sent per country, this is how statistics are normally presented. But if you also take the number of available IP addresses per country it helps in gaining a clearer picture about the spam abuse for each country.

Country	**Number of listings**	**Available IP addresses**	**Percentage**
Brazil	1,398,183	31.852 million	4.38%
India	959,951	18.743 million	5.12%
Russian Federation	682,924	25.635 million	2.66%
Turkey	588,814	10.600 million	5.55%
Poland	488,879	14.066 million	3.47%
Vietnam	339,043	6.711 million	5.05%
China	298,002	204.892 million	0.14%
Italy	229,900	33.117 million	0.69%
United States	223,692	1,479.993 million	0.01%
Ukraine	206,021	5.686 million	3.62%
Thailand	187,914	5.129 million	3.66%
Germany	170,805	85.735 million	0.19%
Korea (South)	168,742	72.321 million	0.23%
Argentina	168,706	7.430 million	2.27%
Romania	166,842	9.403 million	1.77%

(Table 1: Top 15 of current known spam issues per country [7] and number of available IP addresses per country[8])

The number of available IP addresses is taken from Regional Internet Registries (AfriNIC, APNIC, ARIN, LACNIC and RIPE NCC) is not the number of IP addresses in use by these countries. Theoretical the number of available IP addresses and the number that is in use can be way of, however the Internet Assigned Numbers Authority (IANA) handles a strict allocation procedure when it comes to allocating IP space to the different RIR's. The high number of IP addresses for the United States can be a exception in this is because this is where the Internet was originated and where large prefixes where allocated in the early days. For other countries it should be a good measurement.

5 Our research

In our research question we said that we wanted to look if network size, setup and use are reflected in network abuse. For our research we could only use publicly available data.

Size: To find out the network size we looked at BGP information that is associated with a reported IP address. We looked at the most specific prefix that is associated with an IP address and not the allocated prefix that has been assigned by a RIR. We did this because the specific prefix mostly represents a special part of a network. Furthermore the prefix is aggregated from a larger prefix. If we look at this larger prefix we would not get the total size of a network because multiple ranges could be allocated. To get the total number of IP addresses we manually searched for allocated ranges in the specific RIR, we only did this for some interesting networks (See section 5.1).

Setup: For network setup we defined two characteristics. The first one is the difference between dial-up and broadband connections and the second is the use of NAT versus end-to-end connection

Use: Regarding the use of a network (i.e. how much network space is being used in a allocated network) we want to perform active probing on a small dataset and see how many IP's are reported as active with for example a ICMP message.

5.1 Data gathering

To investigate spam abuse we need data to analyze. Since there are no known resources available that offer data on spam abuse over a longer period of time, we decided to create our own data set. For our research we mostly used data from the CBL list, this list consist of unique IP addresses that where reported for sending spam. The list contains about 9 million IP addresses at any given time and is continually updated. Between 8 June 2009 00.00 and 19 June 2009 16.00 (GMT+1) we have downloaded this list every hour and put it's data in a database, this gave us 21,337,779 unique IP addresses over almost two weeks. For every IP address we did a lookup on its BGP ASN, prefix and all information that is associated with it like country, registry and allocated date (See table 2 for the complete set).

The folks over at cbl.abuseat.org were kindly enough to offer us a feed to their blacklist. Because the data inside this blacklist is changing every moment and it was not possible for us to track changes in real-time, we created a window that was big enough to track removals and additions to the blacklist, while preserving disk space on our storage system and allowing us to parse the data into our own data set.

For the BGP lookup information we used the "IP to ASN Mapping" service from Team Cymru. For some IP addresses it was not possible to obtain

information via BGP lookups, the cause for this is that BGP advertisements for those addresses were withdrawn. For 15.946 IP addresses we were not able to obtain extra information via BGP lookups.

Another problem we have observed were IP addresses for which we obtained Multiple Origin AS numbers (MOAS). Because the number (27,918) of reported IP addresses that had more than one AS numbers is relatively low compared to the total number of IP addresses we decided not to investigate this any further. A quick look showed that the number of MOAS addresses is low compared to the total number of reported addresses. However in future work (Section 8) more details could potentially reveal interesting information about these networks, but for our data set we consider it negligible.

The blacklist offered by cbl.abuseat.org is a file that contains as list of IP addresses resembling hosts that have been reported as originators of unsolicited e-mail messages. No other information is contained in this list. When an IP address is added to this list, it can either be de-listed by an administrator responsible for the IP subnet to which the address belongs, or it will be delisted automatically after six days (when no further abuse is reported).

In our data set we decided to store the following information:

Data set	
IP address	The IP address reported in the cbl.abuseat.org blacklist
ASN	Autonomous System Number
CIDR	Most specific prefix
Country	Country belonging to the ASN
Registry	Registry which allocated the ASN
Allocated date	Date when the ASN was allocated
Times listed	Record the number of times the IP address is re-added to the list
Timestamp listing	Timestamp of when the IP address was (re) added to the list
Timestamp delisting	Timestamp of when the IP address was removed from the list

(Table 2: Data set lay-out)

The first thing that we did with the data that we collected was to generate some simple top n statistics regarding spam abuse issues. These statistics are much like statistics that are in use today (See section 4), but do show where to look regarding interesting information. We use these statistics at the end of this paper where we propose a grading method to identify abusive networks (Section 6).

Number of listings	Registry
10,024,395	RIPE NCC
5,204,710	LACNIC
4,690,358	APNIC
775,586	ARIN
598,866	AfriNIC
27,918	MOAS (multiple origin ASN)
5,298	N/A (BGP advertisement withdrawn)

(Table 3: Number of listings per registry)

Table 3 lists the number of abusing IP addresses per Regional Internet Registry (RIR). Caution has to be taken when looking at these numbers, since not all RIR's present this data correctly and might even produce incorrect data due to the Early Registration Transfer (ERX) [9] in 2002.

Number of listings	Country
3,240,425	Brazil
1,820,583	Turkey
1,758,421	Russian Federation
1,708,932	India
1,157,523	Poland
797,630	China
725,701	United States
691,388	Vietnam
583,028	Italy
581,612	Germany
552,699	Ukraine
463,104	Argentina
459,766	Colombia
385,388	Spain
370,704	Thailand

(Table 4: Top 15 of listings per country)

Table 4 lists the top 15 listings per country, calculated from information stored in our data set. In section 5.3 a graph shows additions to the CBL list for the period included in our data set. Also, it shows how the adaption of broadband Internet technologies can be related to the amount of listings at the CBL.

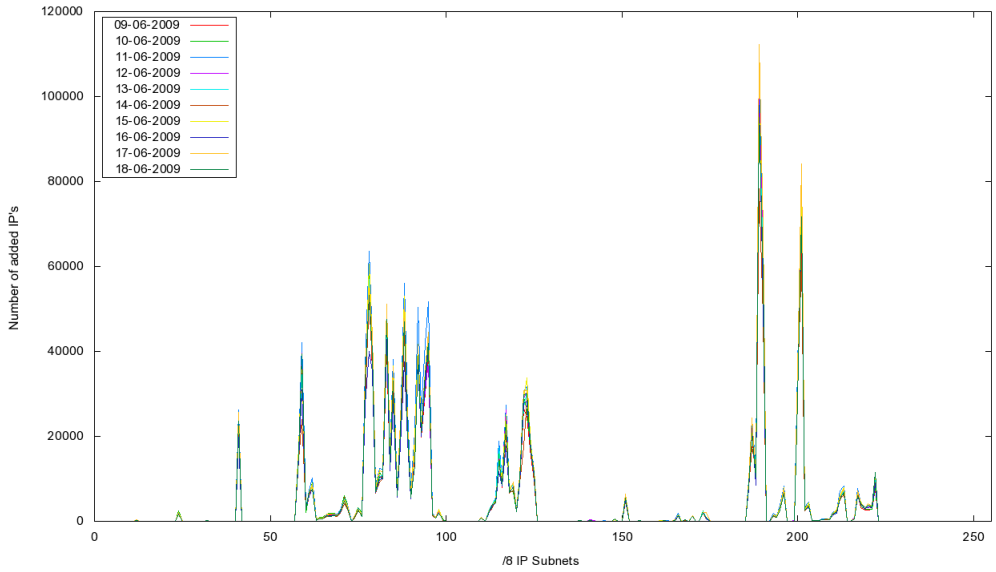
# of listings	Network (ASN - Country - Network name)
1,697,550	AS9121 - TR - TTNET TTnet Autonomous System
984,583	AS7738 - BR - Telecomunicacoes da Bahia S.A.
754,795	AS5617 - PL - TPNET Polish Telecoms commercial IP network
660,910	AS27699 - BR - TELECOMUNICACOES DE SAO PAULO S/A - TELESP
613,119	AS8167 - BR - TELESC - Telecomunicacoes de Santa Catarina SA
602,053	AS9829 - IN - BSNL-NIB National Internet Backbone
453,335	AS7643 - VN - VNN-AS-AP Vietnam Posts and Telecommunications (VNPT)
415,218	AS3269 - EU - ASN-IBSNAZ TELECOM ITALIA
412,337	AS4134 - CN - CHINANET-BACKBONE No.31,Jin-rong Street
374,371	AS24560 - IN - AIRTELBROADBAND-AS-AP Bharti Airtel Ltd
347,480	AS6849 - UA - UKRTELNET JSC UKRTELECOM
287,800	AS7470 - TH - ASIAINFO-AS-AP ASIA INFONET Co.,Ltd
265,931	AS9050 - RO - RTD RTD-ROMTELECOM Autonomous System Number
242,237	AS3320 - DE - DTAG Deutsche Telekom AG
235,307	AS4837 - CN - CHINA169-BACKBONE CNCGROUP China169 Backbone

(Table 5: Top 15 of listings per (ASN) network)

When comparing the country table with the network table some interesting facts become visible, for example you can see that TTNET is responsible for 93.2% of spam abuse issues that are originated from Turkey. Turkey is in the second place in our top 15 spam abuse countries (table 4) but wouldn't be there is it wasn't for TTNET. As with this example you can see that although Turkey is reported as a country where lots of spam abuse is reported it is only because of one network provider. Later we learned that TTNET is actually a big network operator that provides network services for smaller ISP's in Turkey. It is because of 'poor' administration of whois information that all IP addresses from Turkey seem to belong to one ISP.

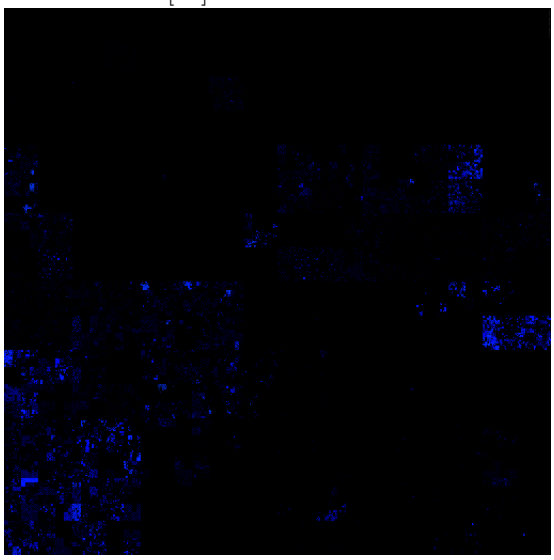
5.2 Distribution of IP listings

Using the collected data of the period between 06-08-2009 and 06-19-2009, we created a graph showing the distribution of IP listings per /8 CIDR block, shown in figure 1. The graph contains separate lines for each day additions were made to the list. In the graph they almost exclusively overlap, we suspect this is due to our limited sample period of two weeks. When looking at the individual graphs for each day, we only observed an increase or a decrease in the number of reported IP addresses.



(Figure 1: IP listings per /8 CIDR block)

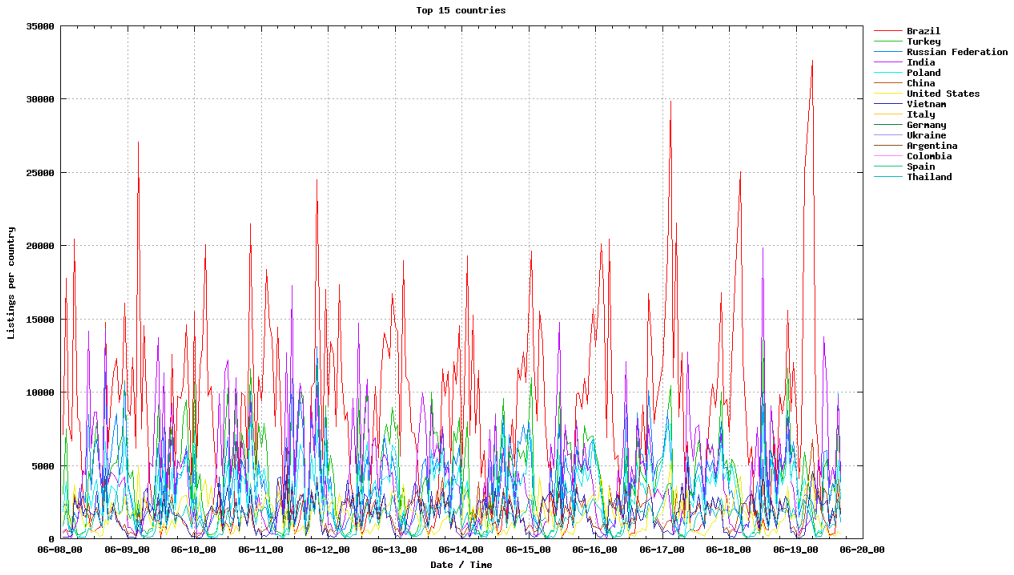
With the use of heatmaps [10] we created an image (See figure 2) that shows one specific time frame in our data set. Interesting to see is that our heapmaps shows great resemblance with other heatmaps that show abuse in broader terms[11].



(Figure 2: CBL Heatmap (09-06-2009))

5.3 Broadband penetration development per country

Figure 3 shows the amount of listed IP addresses per country listed in our top 15 (see table 4). In the graph a day and night rhythm can be seen. Although we did not investigate this extensively, we suspect this is because at night (or after work hours) most computers at offices are switched off.



(Figure 3: Listings of the top 15 countries)

It is interesting to look at broadband as it gives botnet's the ability to generated far higher amount of abuse than with slower dial-up connection. And with the fact that broadband is an "always on" concept a botnet can generate more abuse over a longer time period. Statistics of Internet and broadband penetration are widely available on different Internet resources. The problem is that they only take a small group of countries (i.e. only western world) and rarely over a longer period of time. The statistics that did offer countries that are interesting for our research and offer a longer time period were not up-to-date.

We did find some statistics between 2006 and 2007 (See table 6) that we think are representative for our data set.

Country	2006	2007	Growth percentage
India	15,867	21,107	33%
Russian Federation	10,471	12,707	21%
China	72,408	86,757	20%
Mexico	8,624	10,149	18%
Brazil	12,845	14,964	16%
Italy	15,987	18,106	13%
Canada	18,332	20,392	11%
South Korea	24,297	26,350	8%
Japan	51,450	53,670	4%
France	23,712	24,560	4%
Spain	12,206	12,710	4%
Netherlands	10,772	11,077	3%
Germany	31,209	32,192	3%
United States	150,897	153,447	2%
United Kingdom	29,773	30,072	1%

(Table 6: Internet penetration between 2006 and 2007 (source: opzoeken))

When looking at table 6 it becomes clear that a relation exist between Internet penetration and the amount of abuse (see table 4). The same goes for broadband penetration in table 7.

Country	2003	2004	2005	2006
Turkey	25,531	195,726	506,452	1,530,000
Poland	297,291	818,575	920,752	2,640,000
United States	27,744,352	37,352,520	48,026,587	58,136,577
Italy	2,401,939	4,701,252	6,896,696	8,638,873
Germany	4,513,200	6,904,683	10,706,600	14,085,232
Spain	2,207,008	3,441,630	4,994,274	6,654,881
Netherlands	1,913,200	3,085,561	4,114,573	5,192,200

(Table 7: Broadband penetration between 2003, 2004, 2005 and 2006(source: OECD Stat Extracts))

Broadband penetration, The FCC defines 'broadband' as 200 kbps (in at least one direction), listed in table 7 shows that upcoming countries are responsible for a high amount of spam abuse. This could be because these networks did not need to handle abuse related issues when running smaller and/or slower connections. Some countries listed in table 4 are not in table 7, this is because the source used was the only publicly available source that had the most countries over time that we reported in table 4. However, we found that countries that currently have high abuse numbers, are also high on the list when it comes to broadband penetration growth.

5.4 ISP performance

For comparison we gathered the number of listings for a group of ISP's from the Netherlands. We chose to investigate these ISP's because of our knowledge about their configurations and practices (port 25 blocking, abuse policies). The scores are listed in table 8.

ISP Name	ASN	Total listings	Total unique listings	Customers
KPN	286	1346	1193	1,145,000
SURFnet	1103	145	136	
XS4ALL	3265	1526	1411	289,000
UPC	6830	41455	38278	9,416,700
Ziggo	9143	6412	6045	1,400,000
Online	5390	268	242	335,000
Telfort	5615	2446	2158	438,000

(Table 8: Dutch ISP Performance)

KPN

KPN is, with about 1,145,000 customers[12], one of the largest ISP's in the Netherlands. Most listings in CBL are IP addresses belonging to customers (mostly small and medium-sized enterprises) that subscribed to 'KPN Zakelijk Internet'. As part of this arrangement customers receive multiple public IP addresses. While investigating the listed IP addresses we found two distinct types of listings: first there are listings from multiple IP addresses from the same subnet (from the same customer), and second we found IP addresses that show up more than once in the list. For example the block 193.172.42.0/24 has 26 listed IP addresses in CBL, but when we tried to establish a SMTP connection to one of these addresses no connection could me made. This could be a sign of a malware outbreak at this specific site, abusing the hosts directly connected to the Internet for sending unsolicited e-mail messages. The IP address 193.173.69.190 has been listed (and de-listed) 7 times to the CBL. At this address we managed to establish a SMTP connection. The fact this address has been listed 7 times in only two weeks may suggest this SMTP is configured for open relay or had been compromised by an attacker.

SURFnet

SURFnet is an organization that offers Internet connectivity for higher education and research in the Netherlands[16]. There are not many listings on the CBL and reappearing addresses were not found in our data.

XS4ALL

All XS4ALL ADSL customers, around 289,000[12], are assigned one static IP address from XS4ALL[17]. Via whois information publicly available, distinctions can be made to identify the various services for home users and small and medium-sized enterprises (bdsl). The distribution of listed IP addresses belonging to the XS4ALL network is widespread across the IP space. We did not find specific problem subnets belonging to one customer or IP addresses appearing more than twice on the CBL. Since XS4ALL has a strict policy for customers sending unsolicited e-mail messages[18] abused Internet connections are identified quickly and customers are notified about the abuse and are encouraged to fix vulnerabilities. We think that for this reason most listed IP addresses appear only once on the CBL.

UPC

UPC is a provider operating in several different countries in Europe, delivering around 9,416,700 broadband connections[13]. As one of the few they allow SMTP traffic outside their network. The large number of listings (compared to other dutch ISP's) might be a result of this. The UPC Autonomous System Number is used for several countries where UPC is active, IP addresses that come from the Netherlands were obtained by checking the country field via whois information. Because of this it might be that some listings were missed or some addresses are actually not located in the Netherlands at all. We did not manage to locate specific problem subnets or reappearing addresses from UPC's IP space in the CBL.

Ziggo

Dutch ISP Ziggo is the result of the merger between Multikabel, @Home Network, and Casema[19], having a total of 1,400,000 customers[19]. From the information publicly available it is not possible to determine to which specific service the IP address belongs (like KPN gives detailed descriptions for each subnet). Also the fact that a lot of the whois information available show obsolete data does not help in investigating the listed addresses. It is Ziggo's policy to prohibit SMTP traffic to servers other than the ones they offer for customer use, but since a lot of the listed addresses 'appear' to belong to home customers, we guess this policy is not properly implemented at all parts of the merged networks.

Online

Online is a dutch ISP and is the result of the merger between Wanadoo and Orange Breedband having about 335,000 customers[21]. The provider has very few IP addresses listed on the CBL. Via Online it is only possible to establish a SMTP connection to the servers Online has designated. Also, Online maintains a strict anti abuse policy like XS4ALL does[21]. Online does not offer services

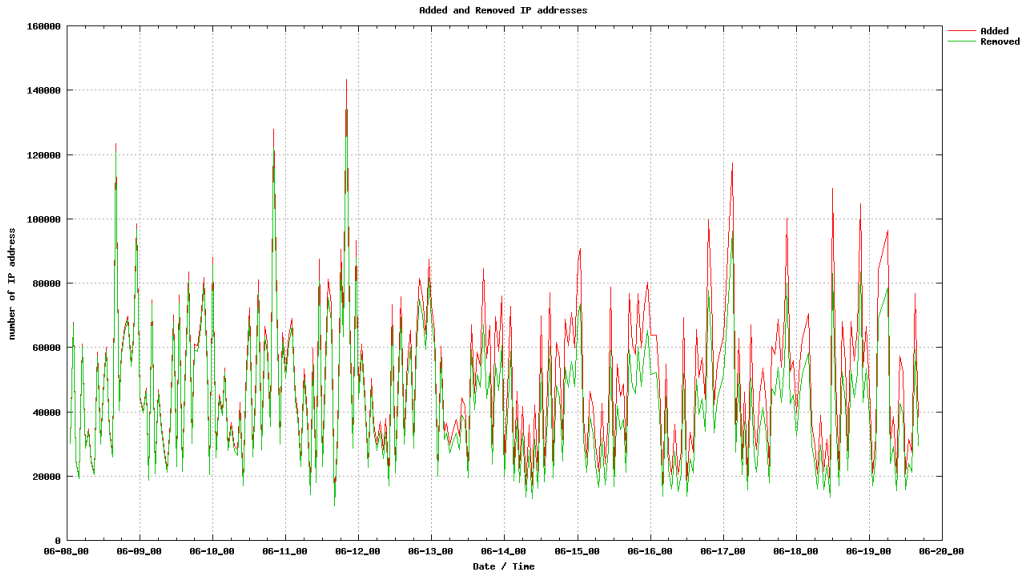
for small and medium-sized enterprises. This is also an explanation for the low number of listed IP addresses. The addresses that are listed on the CBL are hard to describe, since publicly available whois information does not show detailed description for these addresses.

Telfort

Telfort is a dutch ISP, formerly Tiscali, owned by KPN, serving about 438,000 broadband customers in the Netherlands[12]. Like we observed at listings from the KPN IP space, the Telfort listings shares one characteristic: specific /24 subnets with multiple listings are commonly found in the CBL (for example 195.241.197.0/24). Difference is that whois info publicly available is not as specific as for KPN's IP space.

5.5 Abuse development over time

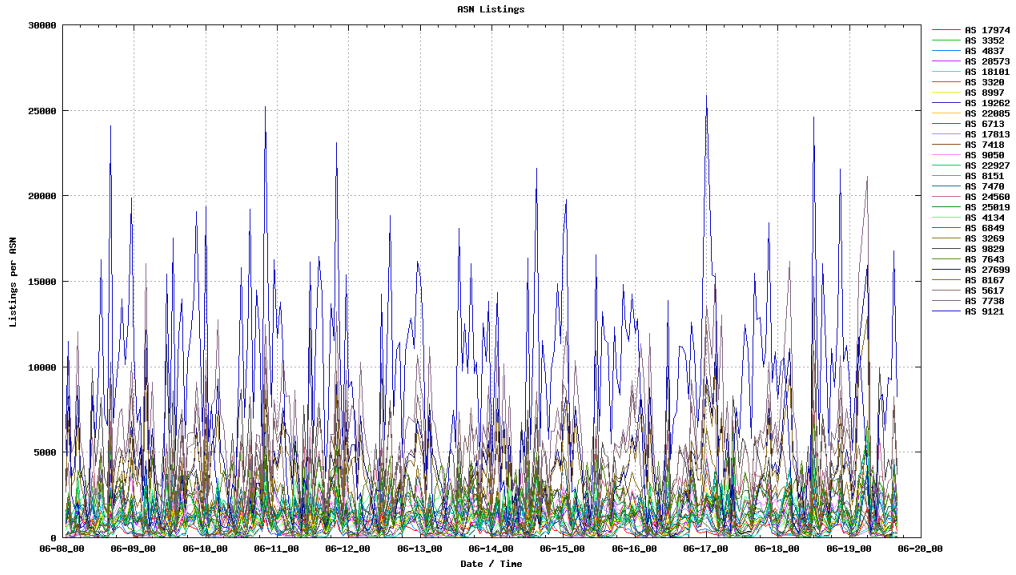
Although the period for which we have gathered data is relatively short, we tried to visualize the distribution of abusive networks in an effort to detect if any major shifts had taken place. With the use of gnuplot we created the following images. These images visualize additions to the cbl.abuseat.org blacklist. The graph shows that additions to the blacklist per day happen for the same /8 IP subnets.



(Figure 4: Listings and Delistings)

Unfortunately we only had time to collect data over a two week period and this didn't gave us a data set when outbreaks or shutdown of spam networks

occurred. However as you can see in figure 4 you can see a clear difference between day and night.



(Figure 5: Listings per ASN, top 10 every hour)

When you look at the the number of listings per network (See figure 5) you don't see any major shift in the number of listings between networks. The only thing that is visible is the number of listings between day and night.

5.6 Network Address Translation

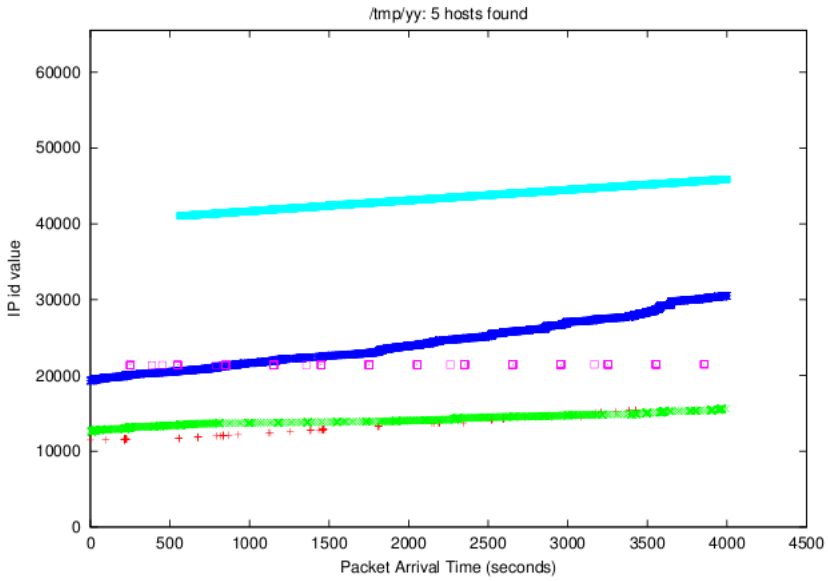
A collection of hosts can share an Internet connection by using techniques like Network Address Translation (NAT)[22]. The use of NAT can potentially influence statistics, since a single IP address can represents multiple hosts. There are several methods available[23][24] for detecting NAT solutions from outside the network, but they only detect if NAT is in place and can't determine the number of hosts. The counting of hosts in a NAT environments is essential if you want to place them in statistics. There are methods available, but they require that you have administrative control over the network that the NAT solution is using. Therefore we will only look at NAT detection in correlation with abusive IP addresses from a theoretical perspective and will leave it for further research.

Not only NAT hosts can conflict with statistics, but also networks with dynamic IP addresses configured with short lease times. If an infected host changes frequently from IP address, it can list multiple IP addresses in a subnet over time. Another cause for multiple addresses getting listed, caused by a

single host, is mobility. Mobile hosts, like laptop's, can be connected to different networks, that can then get listed if the host is sending spam. Like NAT, it is hard to take these realities into statistics without insight knowledge about a network, and therefore we did not take it any further than theory, but it is definitely something to remember.

IP Identification field

The paper "A Technique for Counting NATted Hosts"[25] by Steven M. Bellovin from AT&T Labs Research is the only known method that can count the number of hosts that are using NAT to connect to the Internet. This method primarily uses the "IP Identification field "[26] to determine the number of hosts from a flow of network packets. This identification field is used to distinguish network packets. Most operation systems have implemented this identification as a simple counter, so if you can capture a network traffic flow you could theoretically see these identification fields as an increasing line (See figure 6).



(Figure 6: IPid field from three different hosts (source: AT&T Labs Research [25])

This technique can fail, because some operation systems (i.e *BSD) have implemented the IP identification field as a pseudo-random number rather than a counter.

Analysis of received headers

To determine the percent of spam e-mail sent via hosts that are placed behind NAT devices one could look at the way Mail User Agent's (MUA) identify to SMTP servers with HELO or EHLO commands. Requirements to these

commands, given in RFC 2821[27], state that the MUA has to indicate its identity by issuing the EHLO or HELO command appended with the fully qualified domain name (FQDN) of the client. If a FQDN is not available, then an address literal needs to be given in the form of four small integers separated by dots: [145.100.104.24].

Since most client's do not have a FQDN available to them, the address literal needs to be send for the HELO or EHLO command. When the client is a host behind a NAT device, chances are high it has an IP address as described in RFC 1918[28], and uses this in the address literal sent together with the HELO or EHLO command.

The identity of the client as received by the SMTP server is used to create the Received header for the e-mail message. By looking at these headers in spam e-mail messages, it should be possible to get an estimation on the share hosts behind NAT devices have on overall spam abuse.

A warning with this method needs to be given as well. Since there are no requirements specified in a RFC regarding the HELO or EHLO command and the use of RFC 1918 address space, different MUA's implement different techniques for handling address literals. Some try to determine the public IP address, others like Microsoft's Outlook use the Netbios name. Because of this, received headers can become polluted, and not all spam originating from hosts behind NAT devices can be found.

5.7 Comparison to other abuse lists

There are many different abuse blacklists available[29] on the Internet to stop different kinds of abuse. While this document is mostly based on the CBL spam abuse list, we have looked at other lists for comparison. One of our criteria is that access to the complete list has to be publicly available without any subscripsion or donation. While most abuse blacklists offer their service free of charge by allowing MTA's to do lookups, most of them do not offer the complete blacklist for download. The lists found that matched our criteria are: NJABL[6], SORBS[30] and UCEPROTECT-Network[32]. The NJABL (Not Just Another Bogus List) list has only a couple of dozen new entries added to its list each day. The SORBS (Spam and Open-Relay Blocking System) was closed during our project due to termination of their hosting infrastructure. The only list that we could compare to the CBL lists was the UCEPROTECT-Network abuse lists.

# of listings	Network (ASN - Country - Network name)
156,171	AS9121 - TR - TTNET TTnet Autonomous System
114,238	AS9829 - IN - BSNL-NIB National Internet Backbone
101,146	AS7643 - VN - VNN-AS-AP Vietnam Posts and Telecommunications (VNPT)
99,423	AS7738 - BR - Telecomunicacoes da Bahia S.A.
77,304	AS5617 - PL - TPNET Polish Telecoms commercial IP network
69,466	AS27699 - BR - TELECOMUNICACOES DE SAO PAULO S/A - TELESP
67,496	AS8167 - BR - TELESC - Telecomunicacoes de Santa Catarina SA
57,091	AS24560 - IN - AIRTELBROADBAND-AS-AP Bharti Airtel Ltd
50,883	AS6849 - UA - UKRTELNET JSC UKRTELECOM,
48,549	AS6713 - MA - IAM-AS

(Table 9: Top 10 of listings per (ASN) network) (source: UCEPROTECT-Network[33])

Table 9 lists the top spam abusing autonomous systems, it has much resemblance to the list we created and depicted in table 5.

6 Scoring Networks

With the data we have collected we want to see if it is possible to develop methods that allow us to score the abusiveness of a network, and perhaps provide boundaries on blocking escalation (whole /24 subnets instead of singe IP addresses).

6.1 Growth checking

The first method we describe is one where we take the number of abusive hosts in a /24 subnet and track per day (could be per hour) the amount of listed hosts in this subnet. If for example the number steadily increases from 10 listed addresses on day one to 30 (or more) listed addresses on day 3 we decide to block the whole /24 subnet.

To test how this approach would work we used the data we gathered, identified networks that show the increase in listed IP addresses and determine how many future listings were caught after day 7. We identified 7 suitable networks and their results are depicted in table 10.

/24 Subnet	**Day 1**	**Day 3**	**Day 7**
4.152.210.0/24	14	1	0
4.249.174.0/24	11	9	9
116.21.234.0/24	139	141	151
93.124.44.0/24	9	14	10
201.42.37.0/24	9	9	6
203.81.221.0/24	14	14	17
212.200.116.0/24	24	32	12

(Table 10: Growth Checking)

Unfortunately we did not find any suitable subnets that matched our formulated criteria before escalating to complete /24 subnet blacklisting. There can be a couple of reasons for that, the first one is that maybe our criteria are too low. Network 116.21.234.0/24 shows a steady increase in the number of listed addresses, maybe this increase is a characteristic of networks that have more than one hundred listed addresses. Another reason might be that we did not use enough sample networks. We did apply this method to only seven networks in order to save time, to fully test it all potentially matching networks need to be tested. Due to limited time available to carry out experiments we minimized the sample group.

6.2 Scoring

The second method we propose is one where we take the IP address of a sending host and try to predict the likeliness of a host being a member of criminal classes. For this we look at the historic data we collected and check the following

parameters:

- **Times listed** If the address has been listed before, it receives 1 point for every listing.
- **/24 Score** For every address belonging to the same /24 subnet found in the CBL we add 1 point
- **Route Prefix** (Skip if it is /24) For each /24 in the prefix found in CBL we add 1 point.

The reason we chose these parameters are because they tell something about the host and the network it belongs to. An IP address that has a large history of CBL listings suggests continuing problems at the host the address represents. A subnet that has a lot of listed addresses on the CBL suggest a problem at the network the host belongs to. The same goes for the route prefix and AS number.

It is important to note that the parameter set could be extended by including information like the purpose of the subnet an address belongs to. For example, a description (in whois records) telling the address is part of a range for DSL customers suggest in most situations that e-mail messages are not supposed to originate from this network. Unfortunately we did not have this data available to us. Providers like XS4ALL and KPN keep this information in their whois records at Ripe by using the description field.

Another parameter that could be included is the AS performance over time. For example if the total number of listings in an AS has increased by a number X over the last two weeks, points could be added to the overall score. When the total number has decreased, 'good' points could be scored.

To test this method we took 10 IP addresses and applied the scoring scheme to them, we made use of the CBL of 06/19/200901-00.

IP Address	**Listed**	**/24 Score**	**Prefix**	**Total**
82.184.164.49	12	1	8	21
145.100.104.24	0	0	1	1
82.95.26.112	2	3	348	353
189.90.53.227	2	40	3	45
87.10.99.130	1	12	550	563
89.204.91.208	2	58	83	143
190.209.35.82	2	18	112	132
93.120.190.225	2	92	64	158
77.254.233.195	2	55	686	743
115.241.246.104	1	171	213	385

(Table 11: Scoring IP's)

It can be argued that the scoring per prefix is not justifiable, since a /14 subnet is likely to score higher than a /22 subnet. So perhaps the scoring for the prefix has to be stretched or shrunk to a /16 subnet to obtain more balanced scores. Another way of improving this is taking the score for the number of listings in the /24 subnet, and take that score as an average percentage for each found /24 subnet in the prefixes. This could, however, lead to highly unreliable figures.

By scoring (or perhaps more appropriate, assigning a penalty to) each parameter one can decide to terminate the connection with the sending host after it exceeds a specified threshold.

6.3 Evaluation

The methods described above both try to determine the probability of unsolicited e-mail messages being sent from an IP address. The first method needs lesser resources to operate and is an effective method to prevent unsolicited e-mail being received from networks where, for example, a malware outbreak has taken place. Downside to this method is that legitimate IP addresses can get listed, while never having sent one unsolicited e-mail message.

The second method is more specific because it considers the history recorded for the sending IP address. Combining it with what is known about the subnet it belongs to, and how the network (AS or prefix) operates can give a good approximation about the intentions of the sending host.

A remark needs to be made on listings per /24 subnets. There are ISP's using short lease timers on their DHCP servers, which can lead to situation where multiple listings for a /24 subnet are actually caused by one host. This, however seems not to be common practice for most ISP's, so we believe this does not pollute the CBL data.

Both methods will without a doubt benefit very much from a description field in whois information, that tells if e-mail messages are supposed to originate from a certain IP range. Unfortunately registering information like this, is not common practice for most ISP's.

7 Findings

The goal of this research was to gain better insight in spam abuse in networks. By investigating the data we collected, we found that there are relationships between the characteristics of a network and the spam abuse that takes place in it. We found that mitigating approaches like smtp port blocking and enforcement of strict anti-spam policies can have a positive impact on abuse in networks. Networks that seem notorious spammers, do not seem to enforce these mitigation approaches. The role of size in a network was somewhat difficult to investigate. For example TTNET is a Turkish ISP that offers network services to different smaller ISP's operating on the infrastructure of TTNET. We did manage to find a relation between broadband penetration per country and spam abuse. We suspect that in countries high on the list, networks need to 'mature' to a state where abuse mitigation is a fundamental part of their configuration. Due to limited time we did not manage to perform probing on networks to determine network size and reflect this into the different statistics. We did however looked up some statistics about customer numbers.

By recording history of abusive IP addresses and laying links between them, we found that a lot of useful information on networks can be found in terms of their abusiveness. Multiple listings for an IP address in short time, multiple listings in a subnet or prefix, the time periods between listing and de-listing, all this information can be used to tell something about hosts and the networks they belong to. For this information that we obtained, we tried to develop methods that can be used to tell something about an IP address in terms of its abusiveness. Unfortunately due to limited time available for this research, we did not manage to test the methods extensively.

Although we did not manage to carry out experiments to research the influence of IP masquerading techniques (NAT), we described two theoretical approaches that can be taken to detect the use of NAT. We suspect that these can be used to obtain an approximation on the use of NAT by abusive hosts, but further research is needed.

8 Further research

Throughout this project we came with some ideas that can be done in further research. We only had a limited time window for our research, and only collected data for two weeks. In further research this time window can be extended to see how networks perform over months or year(s). Also extending the sources to multiple blacklists will probably lead to new insights.

BGP advertisement withdraws:
During our research some network blocks that were listed with reported IP addresses withdraw there BGP announcement for a particular prefix, see section 5.1. It could be that this was normal network engineering or maybe there is some correlation with the abuse that was reported. Make sure to also check other prefix withdraws and announces within that ASN and do this over a longer period of time. Maintenance or network issues are sometimes reported on publicly accessible sites, that can reveal the reason for network changes.

Abuse and legislation:
Some countries, or states, are adopting laws to make sending spam illegal. With longer history records it can be determined if, and if: how, the enforcement of these laws had any effect.

NATed host counting and abuse:
In section 5.6 we only showed a theoretical approach for detecting and counting NATed hosts. If one has administrative control over a larger network, this theory can be put to practice. You do not have to collect IP identification fields from all connections, only reported IP addresses will be enough. Remember that the more hosts using NAT, the more difficult it will become to count them.

Personal blocking list:
Some botnet's will only target a specific network. A personal abuse list could prevent this. This concept is already put to use by DShield [34], where it will log and report a hacking attempts with other users to build a personal blocking list. The same could be done for spam abuse. Part of the spam being sent is sometimes targeted to specific domains, which could be used to identify hosts that are more likely to spam a domain.

9 Division of work

- **Michiel Timmers**:
  - Writing first draft project proposal
  - Configuring local ASN lookup deamon
  - Creating parsing scripts for ASN information
  - Investigate top N networks from our data
  - Investigate broadband penetration per country
  - Comparison of CBL data vs other abuse blacklists
  - Creating graphs: Figure 3, 4 and 5
  - Presenting Reasearch statistics, Internet penetration, Grading and Findings
  - Writing report, primarily chapters 4, 5.1, 5.3, 5.5, 5.6, 5.7 and 8
  - Editing final version report
- **Arthur van Kleef**:
  - Editing final version project proposal
  - Parse & store CBL data into our own data set
  - Investigate performance of dutch ISP's
  - Investigate NAT detection by received headers (theoretical)
  - Develop methods to identify abusive networks
  - Creating graphs / heatmap: Figure 1 and 2
  - Presenting Introduction, Current statistics, Resources and Dutch ISP's
  - Writing report, primarily chapters 3, 5.1, 5.2, 5.4, 5.6, 6 and 7
  - Editing final version report

Bibliography

[1] Spamhaus
http://www.spamhaus.org

[2] Universiteit van Amsterdam: System and Network Engineering
http://www.studeren.uva.nl/ma-syst

[3] SURFcert
http://www.surfnet.nl/nl/Thema/surfcert/Pages/Default.aspx

[4] Composite Blocking Lists
http://cbl.abuseat.org

[5] Team Cymru: IP to ASN Mapping
http://www.team-cymru.org/Services/ip-to-asn.html

[6] Not Just Another Bogus List
http://www.njabl.org

[7] CBL: CBL breakdown by country
http://cbl.abuseat.org/country.html

[8] BGP Expert: Number of IP addresses used in every country of the world
http://www.bgpexpert.com/addressespercountry.php

[9] RIPE NCC: Early Registration Transfer (ERX)
http://www.ripe.net/projects/erx/

[10] IPV4 Heatmaps Software
http://maps.measurement-factory.com/software/ipv4-heatmap.1.html

[11] Team Cymru: Hilbert Map
http://www.team-cymru.org/Monitoring/Malevolence/maps.html#hilbert

[12] KPN Q1 2009 Factsheets External Finals
http://www.kpn.com/web/corporate/Corporate-informatie/Investor-Relations/kwartaalcijfers.htm

[13] Q1 09 LGI Press Release Final
http://www.lgi.com/PDF/Q1%2009%20LGI%20Press%20Release%20_Final.pdf

[14] Ziggo Internet Customers
http://www.emerce.nl/nieuws.jsp?id=2736032

[15] Online viert 15 jarig bestaan
http://www.online.nl/?id=375

[16] SURFnet
http://www.surfnet.nl

[17] XS4ALL: Fixed IP address
http://www.xs4all.nl/en/allediensten/toegang/ipadressen/watis.php

[18] XS4ALL: Spam - What XS4ALL does
http://www.xs4all.nl/en/veiligheid/spam/xs4alldoet.php

[19] Ziggo: merger between Multikabel, @Home Network, and Casema
http://en.wikipedia.org/wiki/Ziggo

[20] ZDNet: Overall spam volume unaffected by 3FN/Pricewert's ISP shutdown
http://blogs.zdnet.com/security/?p=3566

[21] Online: Misbruik Melden (Dutch only)
http://www.online.nl/klantenservice/misbruik-melden/

[22] RFC 1631: The IP Network Address Translator (NAT)
http://www.ietf.org/rfc/rfc1631.txt?number=1631

[23] Detecting NAT Devices using sFlow
http://www.sflow.org/detectNAT/

[24] Nessus: Reverse NAT/Intercepting Proxy Detection
http://www.nessus.org/plugins/index.php?view=single&id=31422

[25] Steven M. Bellovin: A Technique for Counting NATted Hosts
http://www.cs.columbia.edu/~smb/papers/fnat.pdf

[26] RFC 791: Internet Protocol Specification (IP ID Field)
http://www.ietf.org/rfc/rfc0791.txt

[27] RFC 2821: Simple Mail Transfer Protocol
http://www.ietf.org/rfc/rfc2821.txt

[28] RFC 1918: Simple Mail Transfer Protocol
AddressAllocationforPrivateInternets

[29] Wikipedia: Comparison of DNS blacklists
http://en.wikipedia.org/wiki/Comparison_of_DNS_blacklists

[30] SORBS (Spam and Open-Relay Blocking System)
`http://www.us.sorbs.net/`

[31] UCEPROTECT-Network Project
`http://www.us.sorbs.net/`

[32] UCEPROTECT-Network Project
`http://www.uceprotect.net/en/`

[33] UCEPROTECT-Network: Current Spammerheavens
`http://www.uceprotect.net/en/rblcheck.php`

[34] Highly Predictive Blacklist
`http://www.dshield.org`