Fonseca *et al. Journal of Internet Services and Applications* (2015) 6:9
DOI 10.1186/s13174-015-0025-5

Journal of Internet Services and Applications
a SpringerOpen Journal

RESEARCH Open Access

Neighborhoods and bands: an analysis of the origins of spam

Osvaldo Fonseca[1*], Elverton Fazzion[1], Pedro Henrique B Las-Casas[1], Dorgival Guedes[1], Wagner Meira Jr[1], Cristine Hoepers[2], Klaus Steding-Jessen[2] and Marcelo HP Chaves[2]

Abstract

Despite the continuous efforts to mitigate spam, the volume of such messages continues to grow and identifying spammers is still a challenge. Spam traffic analysis is an important tool in this context, allowing network administrators to understand the behavior of spammers, both as they obfuscate messages and try to hide inside the network. This work adds to that body of information by analyzing the sources of spam to understand to what extent they explain the traffic observed. Our results show that, in many cases, an Autonomous System (AS) represents an interesting neighborhood to observe, with most ASes falling into four basic types: heavy and light senders, which tend to have many or very few spammer machines respectively, frequent small offenders, where spammer machines appear every now and then but disappear in a short time, and conniving ASes, where most machines do not send spam, but a few are heavy, continuous senders. Not only that, but also by grouping machines based on the campaigns that they send together, we define the notion of *SpamBands*. Those bands identify groups of machines that are probably controlled by the same spammer, and our findings show that they often span multiple ASes. The identification of AS neighborhood types and SpamBands may simplify the combat against spam, focusing efforts at the sources as a whole, possibly improving blacklists by grouping machines found in a same AS or SpamBands.

Keywords: Spam traffic; Autonomous system; Network bad neighborhoods

1 Introduction

In the last two decades, there has been a steady increase in the use of Internet, which led to an increase of the problems related to the sending of spam messages. In addition to the large volume of data generated, since the email service providers estimate that between 40% and 80% of electronic messages are spam, many times they are related to the propagation of phishing [1] and malware [2]. Because of those factors, the losses caused by spam traffic are estimated in billions of dollars [3].

To try to counter those effects, the battle against spammers takes place on several fronts. For example, much has been done to develop filters based on message content, defining rules to identify patterns of obfuscation observed in spam messages [4]. Besides that, in recent years, multiple efforts have focused on understanding spam traffic within the network. The goal in that case is to find elements that can be used to identify the machines that send the messages before they traverse the network and consume resources of mail servers at the destination. This work fits in this line, analyzing where are the machines used by spammers.

Recently, the term Internet Bad Neighborhoods was created to identify contiguous ranges of IP address space that contain a significant number of machines with unwanted behavior [5]. The principle behind the original concept was that machines with similar (bad) behavior that shared an IP prefix would suggest they belong to a same network with problems. Later, the concept was extended to refer to network segments propagating unwanted traffic, regardless of the number of machines involved [6]. The granularity chosen for those analysis was that of /24 address ranges.

Our analysis is based on similar principles, but we chose the level of Autonomous Systems (AS) to identify possible bad neighborhoods. Autonomous Systems, by nature, identify an IP address range under the control

*Correspondence: osvaldo.morais@dcc.ufmg.br
[1] Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
Full list of author information is available at the end of the article

of a single entity responsible for defining usage policies, routing, and administrative procedures to be applied to all machines installed in that range. In this sense, two IP addresses belonging to the same AS would have a much greater chance of exhibiting similar behavior than two IP addresses on subnets with adjacent addresses (the IP space), but belonging to different ASes. After all, an AS with security flaws in its policies is at risk of becoming a potential Bad Neighborhood, since the probability of machines belonging to that AS being infected and starting to send spam is high. It should be pointed, though, that some ASes may be under split management; that could be better treated by considering BGP announced prefixes, but that information was not available in the collected data.

Our data contain spam traffic collected at various points around the globe, which gives us multiple vantage points over the network. By grouping the machines that generate the traffic based on their origin AS, we created a profile of the behavior of each Autonomous System observed during the experiment. Our results show that the majority of machines sending spam are concentrated in a few ASes and that only 15 of them are responsible for over 80% of the observed traffic. By using data mining techniques, we observed that there are similarities between some of them, and that it is possible to group them in four categories representing the AS from the point of view of their role in the distribution of spam. This finding is one of the main contributions of this work, since, until then, it was common to assume that there would be only two types of generators of spam: "light" and "heavy" transmitters. Our characterization indicates that in addition to spam-free and bad neighborhoods (where bad behavior — spamming — is quite common), there are also good neighborhoods, where such behavior is not the norm, but appears from time to time, usually rapidly disappearing shortly after its appearance in one or another machine, and conniving neighborhoods, where misbehaving machines were not widespread, but limited to only a few hosts, which tend, however, to be heavy senders.

2 Related work

This work analyzes spam based on its behavior as seen "inside the network", not based on data from destination mail servers. In that sense, it relates to the works of Ramachandran and Feamster [7], one of the first to study spam from network-level features. Although we consider such features, since we have access to the spam content, we combine different views. Duan, Gopalan and Yuan [8] have recently provided a similar analysis, although based on a single point of observation (a large university campus).

The definition of Bad Neighborhoods, mentioned in the introduction, is due to van Wanrooij and Pras, who proposed the concept as an extension of the use of blacklists in the spam detection [5]. In their study, each 24-bit IP prefix (/24) would be a neighborhood and bad neighborhoods would be those with a large number of machines sending spam. Moreira Moura *et al.* [6] focused on the analysis of these neighborhoods and extended the definition to include IP networks with few transmitters, but with a high volume of traffic, following the classification of "heavy" and "light" transmitters previously proposed by Pathak, Hu and Mao [9]. In our work, we observe that each AS can be seen as a neighborhood, because an autonomous system naturally defines an area with similar machines, since there is a unique management for the whole AS and a common routing policy for all machines.

Many studies analyze spam traffic using messages collected at the destination mail servers. Gomes *et al.* [10] showed features that can be used to separate legitimate messages from spam messages, using data collected from only one specific point of the network. In this work, spam messages were collected by low-interaction honeypots installed in 10 different countries and located in transit networks. This provided a more global view of spam traffic, offering a different perspective.

Kokkodis and Faloutsos [11] showed results that indicate that the activities of botnets are scattered in the IP address space, reducing the effectiveness of anti-spam filters based on addresses and hindering the work of network administrators. Our work, despite confirming the existence of spammers in a very large number of networks, shows that most of the spam messages come from a small number of ASes, a result that can be used in the development of new techniques for spam detection, as in the design of initiatives to act against such sources.

Some of our analysis is based on the concept of spam campaigns. Our definition is based on the identification of frequent patterns in the content, using data mining techniques [12]. Other approaches have been proposed, like the use of regular expressions [13]. Our approach fits better with our processing pipeline, where multiple data mining algorithms are applied to derive different views, such as those in this paper.

This paper is based on previous work, so far available only in Portuguese. In a first paper [14], we performed a detailed analysis of spam messages collected over three months around the world to observe Bad Neighborhoods. With the same dataset, we developed the concept of SpamBands [15], another way to analyze the origin of spammers. (All the major concepts from those papers are included here, to provide a complete source in English). In the current work we extend our analysis of both concepts to cover data from approximately one year, and for the first time we use both concepts, Neighborhoods and SpamBands, to study the relationship between them. That allowed us to identify new patterns, such as the

strong correlation between IP addresses in SpamBands and bad neighborhoods, and a topological relationship among spammers, since the IP addresses from a SpamBand usually come from just a few ASes. We hope that our findings can help drive the spam community's efforts to combat spammers closer to their origin.

3 Methodology

Three aspects of our methodology deserve attention: the collection architecture, and our techniques to identify spam campaigns, and to define *SpamBands*. They are presented next.

3.1 Collection infrastructure

The dataset used in this work was collected using twelve low-interaction honeypots [16] installed in ten different country codes: two in Brazil, two in the United States and one in each of Argentina, Australia, Austria, Ecuador, Netherlands, Norway, Taiwan and Uruguay. That means we had collectors present in four different continents, allowing the study to have a global view of spam traffic. By doing that, we avoided the problem of location bias, which may be present in several studies in the literature, whose data often come from a single collection point. Furthermore, none of the honeypots used in the analysis showed any signs of having been subjected to any form of attack.

The honeypots used in this paper are machines that simulate machines of interest to spammers, such as open SMTP mail relays and HTTP and SOCKS open proxies. Their goal is to lure spammers to identify them as vulnerable machines and use them to try to deliver spam messages. In practice, the honeypots do not deliver spam messages to the intended recipients; instead, they are stored locally and periodically collected to a central storage. The behavior of honeypots, however, is such that it makes the spammer believe that the delivery was successful. That is corroborated by the fact that each most machines continues to abuse the honeypots for all the collection period.

It should be noted that our analysis is guided by the traffic that was directed to our honeypots. There may be spammers that do not make use of proxies/relays to deliver their messages, and those are not considered in this analysis. However, is highly unlikely that a heavy spammer, using a dedicated server farm, would remain in activity without such a technique: it would be easily identified by black lists and blocked, since it uses few origin IP addresses. On the other hand, if a botnet delivers spam directly to the target mail servers all the time, we would not see it in our data.

Along with each message received, additional information is collected and stored by the system. This information includes the protocol used by the spammer to connect to the honeypot, (SMTP, HTTP or SOCKS), the network prefix and AS of origin, the status of the source IP in blacklists like Spamhaus XBL and PBL at the moment each message was delivered, among others. All that is obtained at the time the message is received, so that we have a snapshot of things as they were at the moment the spammer tried to send each message. Thus, our analysis considers the information available at the time of the transmission and not during a later query, which could cause error. That is essential, for example, for the analysis of black list contents, which might change between the time of collection and analysis.

Later, during the analysis, some ASes deserved further study. In those cases, based on their AS numbers, we gathered data available on the Internet to get more details about their activities. Based on the activities that were identified during that search, we classified the ASes as providers (general, DSL, corporative), hosting/co-location services, etc.

3.2 Spam campaigns and spam bands

To better understand the behavior of spammers, we used the concept of spam campaigns. A campaign is a set of messages that share a common goal (similar content) and a common dissemination strategy [12]. We used the FPCluster algorithm to group messages based on their various attributes and to identify the obfuscation strategies used. That algorithm builds a frequent pattern tree, which is then used to extract the message clustering patterns, which in turn identify the campaigns [12,17].

Through the identification of campaigns, we detect the influence of each orchestrated campaign on the spam traffic collected, as well as the emergence of new IP addresses that join a given campaign. Based on those observations, Fazzion *et al.* [15] developed a method that can identify groups of transmitters that are correlated, called *SpamBands*. Since that work was published in Portuguese, the method is described here for completeness.

The premise of *SpamBands* was that machines which generate the same campaigns are controlled by the same orchestrator, being related in terms of dissemination strategy used. Thus, a *SpamBand* is a group of machines that works together on the same set of campaigns. The relationship between machines and campaigns can be modeled as a graph G, where the machines are vertices and there is an edge between two machines if they sent messages associated with the same campaign. Figure 1 illustrates the construction of this graph.

From G, we can define a *SpamBand* as a dense subgraph that can be obtained by several clustering algorithms in graphs in the literature which can be quite complex and hard to calibrate [18]. Our strategy is more simple and interactive. Initially, each *SpamBand* is a connected component.

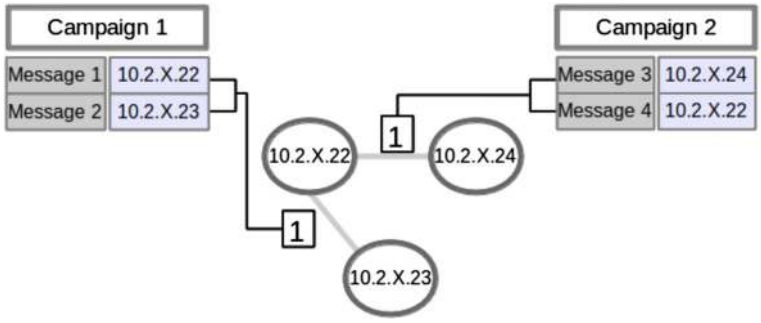

Figure 1 SpamBand graph construction. The graph represents IP addresses that sent spam as vertices. There is an edge between two vertices if they took part in the same spam campaign.

In some cases, however, one IP address may be found connected to more than one such sub-graph. That may be due to IP address reassignment, or use of NAT. To handle that, in a second moment, we evaluate those cases, which can require the split of certain connected components in order to isolate subgraphs with higher density.

The process to identify *SpamBands* is presented in Algorithm 1. The algorithm receives three parameters: the graph (**G**), the minimum threshold betweenness to be considered (**threshold_bt**) and the maximum number of IP addresses that can be removed in order to split a connected component (**threshold_ips**).

Algorithm 1: *SpamBand* (Graph G, Float threshold_bt, Float threshold_ips)

```
S = ∅;
C=G.ConnectedComponents(); ;
for comp in C do
    ips_to_remove = ∅ ;
    for ip in comp do
        if ip.Betweenness() >
        threshold_bt*comp.BiggestBetweenness() then
            ips_to_remove.Add(ip);
        end
    end
    if ips_to_remove.size() >
    threshold_ips*comp.NodesNumber() then
        S += comp;
    end
    else
        S += comp.RemoveNodes(ips_to_remove);
    end
end
return S;
```

The first step is to determine the connected components in G which constitute the initial approximation of the *SpamBands*. Next, we identify dense sub-graphs in each connected component exploring the *betweenness* concept, which measures the centrality degree of a node in the graph. This metric indicates the number of shortest paths among all pair of nodes in the graph that pass through a given node. Our premise is that when some nodes have a high value of *betweenness*, beyond what would be expected for a strongly connected graph, chances are that those nodes are connecting two (or more) sub-graphs which are, themselves, internally dense. Thus, by removing those nodes, we are emphasizing the separation of those internally dense sub-graphs. This removal is based on the parameters **threshold_bt**, which is the lower bound of *betweenness* that a node may have in order to be removed, and **threshold_ips**, which defines a maximum threshold of the number of nodes that can be removed in order to split a component. Algorithm 1 initially verifies which nodes satisfy the *betweenness* threshold in each connected component and next verifies if their removal does not lead very small graphs. If it is possible to remove the nodes, each resultant component is inserted in S. If not, the current component is inserted in S. The algorithm returns the set S which holds all *SpamBands*.

4 Collected data

Our analysis considers approximately one year of collection, from May 9, 2012, until March 31, 2013, resulting in nearly four billion messages (14 TB). By analyzing a large period, we avoid any impact due to an atypical behavior, which could occur in a short period of time.

Table 1 shows an overview of the data collected by the twelve honeypots and used in the study, broken down by the protocol used by the spammers. During the period of almost a year, 3.97 billion messages were collected, which correspond to 14 TB of data. The addresses of the machines that sent spam were associated with 149 different country codes, which corresponded to about 60% of all country codes. We can also notice a large number of autonomous systems, 3,226, showing that the collection included many subnets of origin.

Table 1 Global vision of the data

	SMTP	SOCKS	HTTP	Total
Messages ($x10^6$)	690 (17.4%)	2,486 (62.5%)	799 (20.1%)	3,976
IP addresses	294,072	34,397	11,449	328,050
Autonomous systems	3,096	443	55	3,226
Country codes	146	66	14	149
Volume (GB)	2,564	8,522	3,378	14,464

The number of IP addresses using SOCKS and HTTP protocol is much smaller when compared to the number of IP addresses that used the SMTP protocol. Nevertheless, the number of messages sent using HTTP and SOCKS is larger. This shows that there is not a direct relationship between the number of machines and the number of messages sent. This division is a sign of the differentiation between spammers: some adopt strategies based on high volume over a certain protocol, while others may send lower volumes, using more machines, over another protocol. In fact, during our analysis we will see that there are more factors to be considered.

5 Neighborhood analysis

In this work, we advocate that ASes can be used for the identification of the limits of the neighborhoods, instead of /24 prefixes, as used in the original definition [5]. That provides a more natural aggregation of addresses, given that fixed-length prefixes are not adequate for all cases.

To show that, in Table 1 we observe that spam messages come from many different networks, since 3,226 distinct autonomous systems appeared in the collected data. It is interesting to notice that most of those spam messages were sent by a very small number of autonomous systems, being fifty of them responsible for over 85% of all traffic. Thus, analyzing the behavior of spam at the source can direct the efforts to fight spam as it can identify which are the neighborhoods that have worst behavior, and, consequently, that are more likely to send spam messages.

5.1 Distribution of IP addresses in autonomous systems

Figure 2(a) shows that for the majority of Autonomous Systems, only a few devices were seen contacting the honeypots. In almost 90% of the ASes, the number of machines observed was smaller than 20. In Table 2, we can clearly see the existence of a large group of AS with a small number of IP addresses sending spam messages and another group, smaller, which contains most of those addresses. This shows that the machines that send spam messages are not evenly distributed across the Autonomous Systems.

Those ASes that have fewer than 10 machines that send spam, account for over 83% of the total. Nevertheless, they send only 7.66% of the messages and correspond in number of machines to 1.7% of the total. Thus, we believe that in terms of neighborhoods, these AS are not characterized as bad neighborhoods, but that their security policies are being implemented correctly, because of the small number of spamming IP addresses and the low volume of traffic generated by them. On the other hand, 95 autonomous systems (2.94%) have more than 319,000 IP addresses in the dataset (97.41%) and are responsible for 71% of traffic from spam, which corresponds to almost 3 billion messages. Those neighborhoods show bad behavior, possibly due to weak security policies. Thus, direct efforts to understand and improve the behavior of those networks might have impact on the overall traffic.

Figure 2 also shows the distribution of IP addresses present in each blacklist. The two black lists considered here are XBL, which lists IP addresses detected as infected, and PBL, which lists IP ranges declared by ISPs as being used for dynamic hosts — which should not send mail directly. Finally, we consider IPs that were not found in any of the blacklists considered (No BL). There is a very small number of Autonomous Systems that do not have IP addresses in XBL, about 15%, as we can see in Figure 2(b). In addition, approximately 60% of the ASes have all their IP addresses in XBL. This result makes us believe that most IP addresses are detected by the XBL, but what happens is that a good portion of the ASes listed there (49%) have only one spamming IP address (therefore, 100% of their addresses are in the XBL).

Considering that, the information about the IP addresses that are in the PBL and those who do not participate in any blacklist (nobl) end up getting distorted, as shown in Figures 2 (c) and 2 (d). In the graph of Figure 2 (c), for example, more than 70% of autonomous systems have no IP address present in the PBL, but more than 85% of IP addresses found in the period of analysis are in PBL as shown in Table 3. The IP addresses that are not in any blacklist, belong mostly to a few ASes, and represent a very small portion of the IP addresses, less than 12%. However, this small number of machines is responsible for over 68% of all spam traffic, as shown in Table 3.

5.2 Analysis of neighborhoods with higher transmission power

Table 4 shows the 15 ASes that sent the most spam messages in the period analyzed, being, by themselves, responsible for more than 80% of all spam traffic. This information can be incorporated into spam filters and used by network administrators to reduce the volume of spam, since messages sent by those AS tend to be spam. By analyzing them, in spite of the fact that all send a large volume of spam, we see that there are some neighborhoods

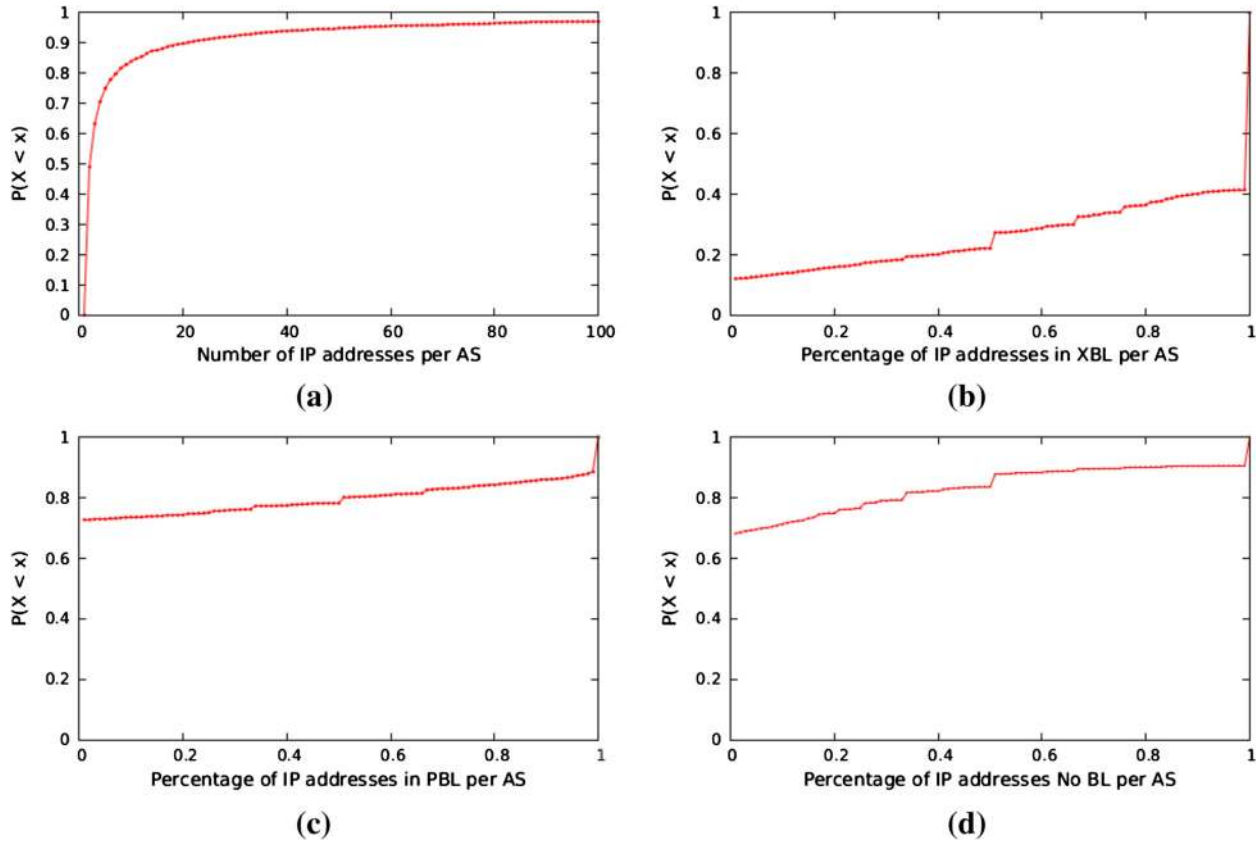

Figure 2 Distribution of active IP numbers per AS, as well as the distribution of the number of IPs from an AS present in the XBL, PBL black lists (or not present in any BL). **(a)** CDF of the number of IP addresses per AS. **(b)** CDF of the percentage of IP addresses in XBL per AS. **(c)** CDF of the percentage of IP addresses in PBL per AS. **(d)** CDF of the percentage of IP addresses that are not in any blacklist per AS.

that have very different characteristics from each other. On the other hand, it is possible to notice that some of them are very similar, although they have no direct relationship.

Some autonomous systems (10297 and 29802) have similar characteristics in virtually all aspects. They have a small number of IP addresses in our dataset, most of them using SOCKS and HTTP protocols to send spam, and do not belong to any blacklist. AS 2497 is also very similar, despite having a larger number of machines. AS 4725, in turn, differs only by having a large number of IP addresses in PBL blacklist. The machines of those neighborhoods behave like dedicated servers used to send spam: they use SOCKS and HTTP protocols, meaning they do not contact any mail host directly (only through proxies), each sends a large number of messages, and most of them are not in any blacklist.

In contrast, we find some neighborhoods with completely different characteristics, such as Autonomous Systems 3462 and 4134. Both have more than 100,000 IP addresses in our dataset, the vast majority of machines observed used the SMTP protocol to send spam and most of them were in some blacklist. In addition, AS 4134 has very striking features, with more than 99% of their IP addresses sending spam messages using the SMTP protocol and about 17,000 of them in XBL.

Table 2 Number of IP addresses observed per AS

IP addr. per AS (x)	#AS	#Msgs (x10^6)	#IP addr.
x = 1	1,581 (49.0%)	108 (2.7%)	1,581 (0.5%)
x <10	2,705 (83.9%)	305 (7.7%)	5,635 (1.7%)
x <50	3,061 (94.9%)	797 (20.0%)	13,309 (4.6%)
x <100	3,131 (97.1%)	1,150 (28.9%)	18,159 (5.5%)
x ≥ 100	95 (2.9%)	2,825 (71.1%)	319,554 (97.4%)

Table 3 Overview of blacklists

	#IP addresses	#Messages (10^6)
XBL	66,388 (20.2%)	594 (14.9%)
PBL	282,599 (86.2%)	789 (19.8%)
No BL	41,005 (12.5%)	2,714 (68.3%)

Table 4 15 most important autonomous systems

AS	Msgs (10^6)	IP addr.	IP SMTP	IP SOCKS	IP HTTP	IP XBL	IP PBL	IP No BL	vol. (GB)	Classification
10297	1.857	182	22.5%	77.5%	77.5%	17.0%	0.0%	83.0%	5,475	hosting/co-location
3462	359	100,395	78.7%	22.1%	6.8%	7.7%	99.8%	0.1%	1,320	DSL/ISP
29802	298	25	16.0%	84.0%	84.0%	12.0%	0.0%	88.0%	781	hosting
9299	148	82	39.0%	61.0%	26.8%	34.2%	1.2%	65.9%	342	DSL/business
2497	141	1,185	0.2%	99.6%	99.0%	0.2%	19.4%	80.4%	559	hosting/clouding
4134	126	110,123	99.7%	0.3%	0.2%	15.8%	81.8%	17.1%	2,446	DSL
6648	124	54	9.3%	90.7%	40.7%	9.3%	90.7%	1.9%	279	DSL/business
4725	31	215	0.9%	99.1%	98.6%	0.9%	95.8%	3.3%	120	clouding/business
27699	31	1,382	7.7%	92.3%	0.0%	9.2%	95.2%	2.7%	114	DSL/business
18881	28	3,091	10.7%	89.3%	0.0%	8.6%	96.1%	2.0%	97	co-location/ISP
8167	25	198	38.9%	61.1%	0.0%	37.4%	34.8%	41.4%	95	clouding/ISP
4837	23	20,551	99.9%	0.1%	0.0%	26.5%	78.4%	18.5%	97	-
9924	22	1,959	1.3%	82.1%	98.6%	2.8%	99.6%	0.1%	77	-
28573	21	700	75.7%	24.3%	0.0%	46.9%	98.4%	0.9%	76	ISP
4230	20	429	21.7%	78.3%	78.3%	31.0%	67.6%	11.0%	67	hosting/ISP

5.3 Grouping of autonomous systems

Because of the evidence mentioned in Section 5.2, we looked for a way to group the AS observed and classify them according to their characteristics. For this, we use the X-means clustering algorithm [19], considering the characteristics of each neighborhood as attributes. The algorithm has the quality of automatically setting the optimum number of clusters to use, unlike other clustering algorithms.

To perform the clustering, we used as features the characteristics that better represent the Autonomous Systems in our analysis. The attributes carry information such as number of IP addresses observed, number of messages per day, percentage of the IP addresses in blacklists, percentage of IP addresses using each protocol, and the average number of messages sent per IP address. Those attributes proved to be a good set to identify the neighborhoods, because they define the major elements of behavior that machines on those networks can present.

Table 5 exhibits the properties of each of the four groups generated. Group 1 contains 64% of the ASes, and most of them have a very small number of observed IP addresses. Most of the machines in that group use SMTP protocol and are in XBL. Group 2 also has a small number of IP addresses, but it is responsible for most of the spam sent (65%). In addition, more than 98% of the messages were sent using SOCKS and HTTP protocols, although most of the IP addresses (84%) use the SMTP protocol to send spam. The vast majority of messages (97%) was sent by machines that were not in any blacklist. Group 3 differs from the others because more than 99% of its IP addresses sent messages through the SMTP protocol and most of them are in PBL and XBL. Finally, the fourth group contains the ASes with a large number of machines observed in the dataset. The majority of their IP addresses are in PBL, and only a smaller number is in XBL. Although most of the machines used SMTP, the highest volume of spam was sent using the SOCKS protocol.

Table 5 Features of each group

	Group 1	Group 2	Group 3	Group 4
ASes	2,064	449	359	354
Msgs (x10^6)	379	2,602	88	907
No. IP Addresses	8,426	16,024	11,503	301,760
Msgs/IP (x10^3)	45.0	162.4	7.6	3.0
Activity[1]	48.8	63.4	67.1	85.3
Msgs-SMTP	85.3%	1.5%	71.4%	29.2%
Msgs-SOCKS	13.5%	75.3%	23.2%	50.3%
Msgs-HTTP	1.2%	23.2%	5.4%	20.5%
Msgs-Xbl	81.9%	1.7%	68.6%	20.0%
Msgs-Pbl	2.8%	1.4%	52.0%	76.8%
Msgs-No-Bl	17.2%	97.2%	15.0%	11.7%
IPs-Xbl	83.5%	13.2%	64.7%	16.5%
IPs-Pbl	5.2%	12.9%	86.6%	89.5%
IPs-No-Bl	15.2%	77.7%	7.1%	8.8%
Volume (TB)	1.2	7.55	0.36	4.98
IPs-SMTP	97.43%	84.62%	99.19%	89.01%
IPs-SOCKS	2.67%	15.27%	0.82%	11.12%
IPs-HTTP	0.33%	9.52%	0.02%	3.44%

[1] Average of days, in the analyzed period, in which the machines of this group were active.

If we consider the neighborhoods that sent more spam messages, studied in Section 5.2, we see that the clustering placed ASes with similar characteristics in the same group, and separated those with different behaviors. Autonomous Systems 10297, 29802, and 2497, whose machines behave like dedicated servers, ended up in group 2, responsible for most of the spam traffic, even though having fewer IP addresses. Moreover, that group has few machines in blacklists, which is a necessary feature for machines that send a large volume of messages — otherwise they would not be effective.

Most machines from Autonomous Systems 3462 and 4134 behave like bots and those two neighborhoods are part of the Group 4. That group includes ASes that have a large number of IP addresses with most of them in blacklists. It is also observed that most of the IP addresses in that group sent a small amount of messages.

By analyzing the 15 neighborhoods highlighted in Section 5.2, we found that none of them are in groups 1 or 3, as can be seen in Table 4. This result was already expected, since the ASes in group 1 have a very small number of IP addresses and those from group 3 have few machines that send spam and are responsible for few spam messages. Thus, the neighborhoods that send more spam were allocated to the other two groups: ASes that have a lot of IP addresses and those that send a large amount of spam messages.

We believe these results may be used in at least two important ways: to help guide policies used by different network managers in the way they treat data from ASes known to fall into a certain class, and to help the network community to identify organizations that may be in need of some orientation on how to handle their security (those with a large number of low volume spamming IPs), or those that may require some pressure to act against server-heady spammers that may be among their clients.

5.3.1 Group 1

The graph in Figure 3(a) shows that a very small percentage of the neighborhoods of this group use the SOCKS and HTTP protocols to send spam messages and that over 70% of the AS send less than 100 thousand of messages in the period. Although the number of messages is small, given that we collected for almost one year, we can see in Table 5 that this group had a very small number of IP addresses observed, resulting in a relatively high number of messages by IP address.

The main characteristic of the ASes of this group is the small number of spamming machines, as we can see in Figure 3(b). Almost 60% of the Autonomous Systems here have only a single IP address in the dataset and none of them have more than one hundred IP addresses. This explains why that group, even encompassing 64% of the AS, is responsible for only 9.5% of the spam traffic generated. Furthermore, most of the IP addresses in this group are in XBL blacklist, which characterizes infected machines, probably belonging to botnets.

The activity period of the machines in the AS of this group show constant activity using the SMTP protocol in almost all ASes, as shown in Figure 4(a). On the other hand, the graph of Figure 4(c) shows that few ASes use the HTTP protocol and only one used this protocol for more than ten days. This large amount of messages sent by SMTP protocol, along with a high number of IP addresses in XBL, suggests the existence of bots. As there are few committed IP addresses in these AS, it is possible that those machines constitute exceptions in the security policy of an overall secure system and that, for some reason, go unnoticed to the management of these networks.

5.3.2 Group 2

This group contains the ASes that sent more spam messages and together they are responsible for more than 65%

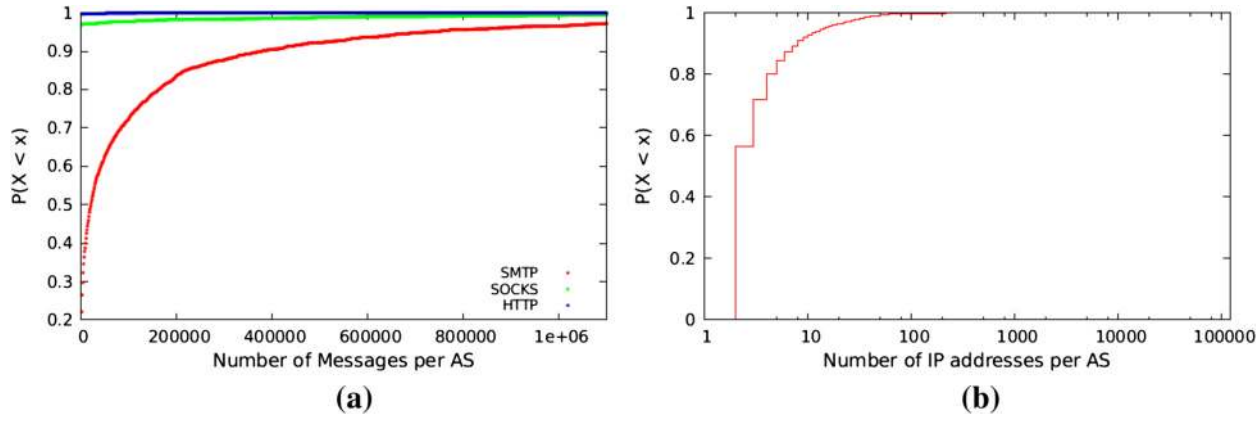

Figure 3 Distribution of messages sent and active IP numbers for ASes in group 1. **(a)** CDF of the number of messages sent per AS. **(b)** CDF of the number of IP addresses per AS.

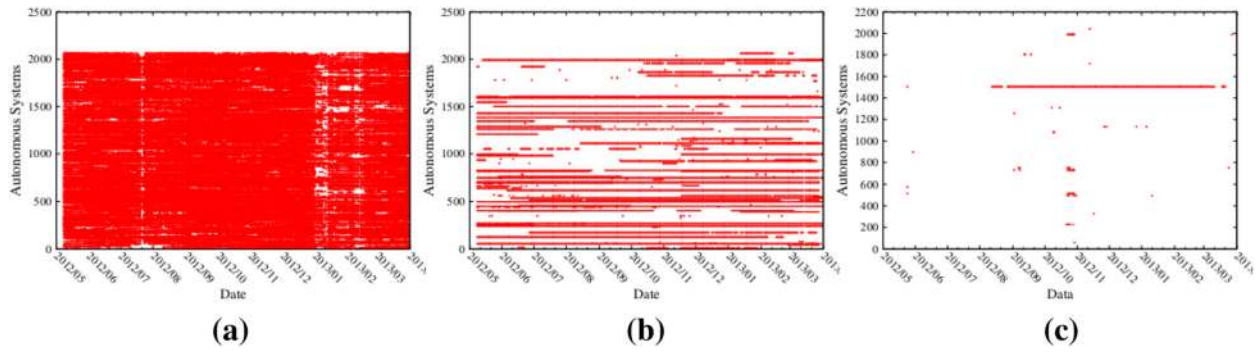

Figure 4 Activity of Autonomous Systems in group 1, per protocol. **(a)** SMTP. **(b)** SOCKS. **(c)** HTTP. ASes appear ordered in the y axis. There is a mark at position (x,y) if AS y was active on date x.

of all messages. As shown in Figure 5(a), 5% of the ASes sent more than one million messages and are responsible for most of the spam traffic. In this group, almost no message is sent using the SMTP protocol, since 98% of the messages were sent by SOCKS and HTTP protocols.

The Autonomous Systems in this group also have a small number of spamming IP addresses — 57% of them had only one IP address in the dataset. Moreover, a very small percentage of neighborhoods here (2%) have more than one hundred machines. However, even with a small number of IP addresses, the average number of spam messages sent by each of them is very large, more than 162 thousand, as can be seen in Table 5. Those features (few machines, with heavy spam traffic) suggest that most of the ASes here house machines that act as dedicated servers to send spam, probably with the connivance of the network administrators. In our opinion, an unwanted bot that would start behaving that way would not go unnoticed by a network administrator that did not accept such practice, and it would not remain limited to a few machines if the network administrator was careless enough not to bother about it. One final interesting aspect is that, in this group, most of the IP addresses are not in any blacklist. Considering the volume of traffic they generate, that would only be possible if they consistently abuse intermediary machines to hide from blacklist detection.

Compared to the other groups, the period of activity of the AS using SOCKS and HTTP protocols are higher in this group. We can see that there is a larger number of ASes and they remain active for a longer period, as the graphics of Figure 6 clearly shows.

As mentioned earlier, ASes 10297, 29802 and 2497 were assigned to this group. Like others in the group, that were studied, those AS are characterized by offering hosting and co-location services, which would fit the profile just described.

5.3.3 Group 3

Table 5 shows that in this group, over 70% of spam messages were sent using the SMTP protocol and the ASes

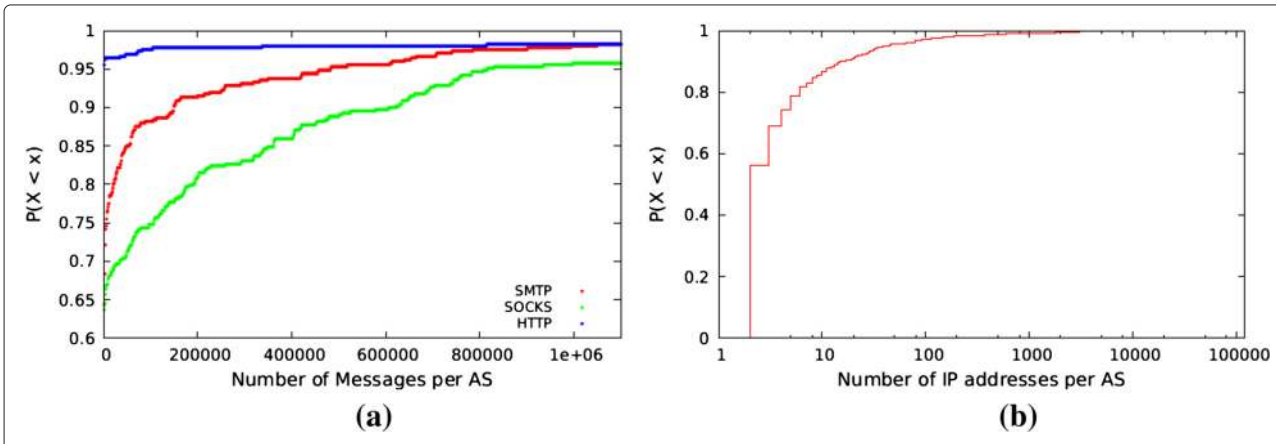

Figure 5 Distribution of messages sent and active IP numbers for ASes in group 2. **(a)** CDF of the number of messages sent per AS. **(b)** CDF of the number of IP addresses per AS.

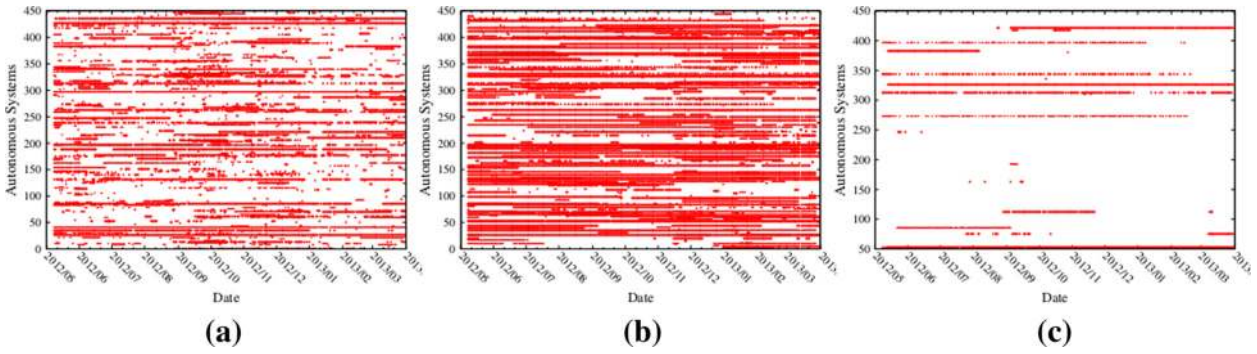

Figure 6 Activity of Autonomous Systems in group 2, per protocol. **(a)** SMTP. **(b)** SOCKS. **(c)** HTTP. ASes appear ordered in the y axis. There is a mark at position (x,y) if AS y was active on date x.

of the group are responsible for only 2.2% of all messages. The graph in Figure 7(a) shows that over 80% of the neighborhoods here sent less than 200,000 messages, explaining the fact that this group is responsible for a smaller number of messages.

The graph in Figure 7(b) shows a similar behavior to that seen in groups 1 and 2, but the number of ASes with only one IP address is lower, just under 40%. What marks this group is the large number of IP addresses that use the SMTP protocol, over 99% of them, surpassing any other group. In addition, about 64% of the machines in this group are in XBL. This suggests the presence of bots, but the low number of IP addresses suggests that there are fewer compromised machines in those AS.

The graph in Figure 8(c) shows that only one AS of this group sent spam messages by HTTP protocol, and just for a short period of time. As expected, for the SMTP protocol, all the ASes were very active throughout the period, as seen in Figure 8(a).

5.3.4 Group 4

From Figure 9(a) we can see that most of the ASes in this group used the SMTP protocol to send spam, but the few Autonomous Systems using SOCKS and HTTP protocols sent more than one million messages. The neighborhoods of this group are responsible for about 23% of all spam traffic.

This group contains the ASes with the larger numbers of machines observed, as can be seen in Figure 9(b), with over 20% of neighborhoods with over 1,000 IP addresses, in which some of them have more than 100,000 machines. Thus, even accounting for much of the spam traffic, the number of spam messages per IP address is the lowest among the groups, only 3,000. Moreover, the great majority of the IP address are in blacklists and use the SMTP protocol. For all this, we have strong evidence that many of the machines belonging to this group are part of botnets. Because of the large number of machines in this situation, these AS are classified of bad neighborhoods, where,

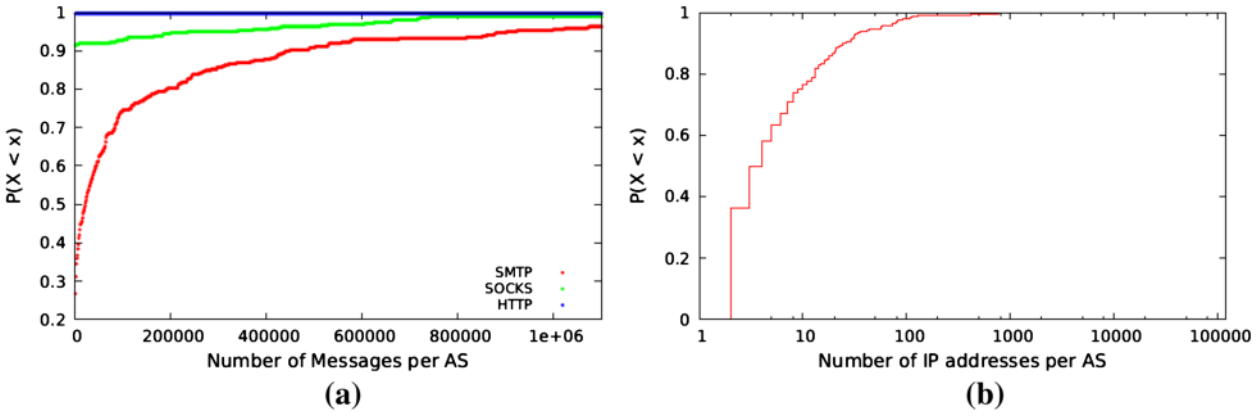

Figure 7 Distribution of messages sent and active IP numbers for ASes in group 3. **(a)** CDF of the number of messages sent per AS. **(b)** CDF of the number of IP addresses per AS.

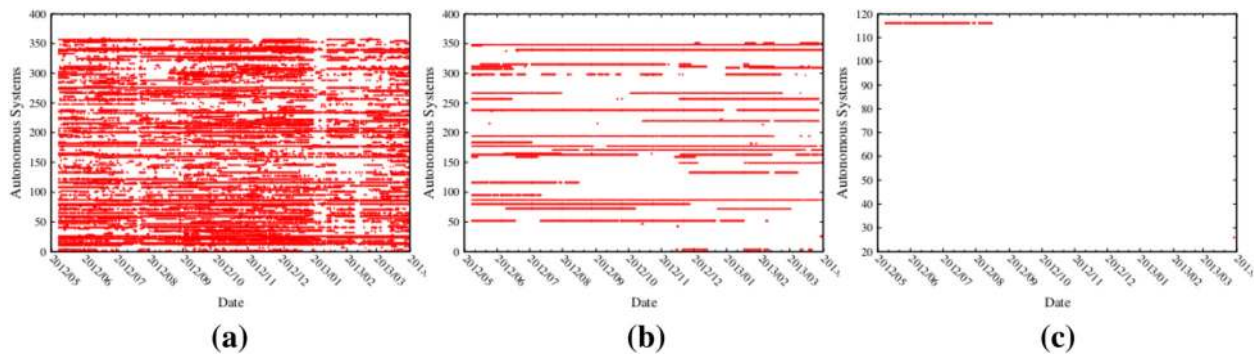

Figure 8 Activity of Autonomous Systems in group 3, per protocol. **(a)** SMTP. **(b)** SOCKS. **(c)** HTTP. ASes appear ordered in the y axis. There is a mark at position (x,y) if AS y was active on date x.

apparently, management policies and network maintenance are not able to prevent the proliferation of infected machines.

Because the ASes in this group have a very large number of IP addresses, it is common for the same AS to show the use of the three different protocols in the dissemination of spam. This behavior is explained by Figures 10(b) and 10(c). It was expected an intense period of activity in the use of the SMTP protocol, once machines belonging to botnets tend to use this protocol. Therefore, the graph of Figure 10(a) reinforces the suspicious about botnets.

ASes 3462 and 4134, which are part of this group, have been classified as ISPs with DSL networks. This suggests that the composition of this group is predominantly domestic users machines infected by some type of malware.

6 *SpamBands* analysis

We applied the *SpamBand* identification algorithm described in Section 3 to the data from each honeypot. We found a total of 2618 *SpamBands* and the distribution of those among the honeypots is shown in Figure 11(a). This Figure reinforces the notion of orchestration by spammers: all honeypots have a well-defined range of *SpamBands* that attacked them each day, and the increase and decrease in the number of *SpamBands* in that interval suggests an orchestration in order to obfuscate the action of groups of spammers.

Figure 11(b) shows a linear regression of the number of *SpamBands* per day for each honeypot. The linear trends reveal lines with low inclination (almost constant) adding to the impression that the variation observed in Figure 11(a) is regular and is due to some kind of obfuscation. Another interesting result is about honeypot EC-01. That honeypot was attacked by more *SpamBands* than any other, although no clear reason for that was found.

Figure 12(a) shows that only 5% of *SpamBands* using SOCKS or HTTP have more than 100 IP addresses, again suggesting the use of dedicated infrastructure for sending messages. In contrast, about 30% of total *SpamBands* using SMTP have more than 100 IP addresses, which is not surprising, since these are supposed to be botnets

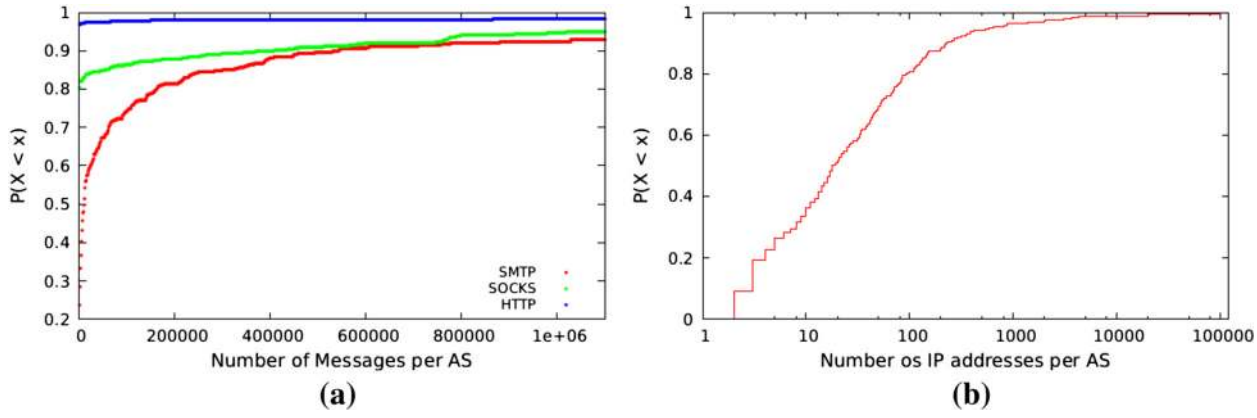

Figure 9 Distribution of messages sent and active IP numbers for ASes in group 4. **(a)** CDF of the number of messages sent per AS. **(b)** CDF of the number of IP addresses per AS.

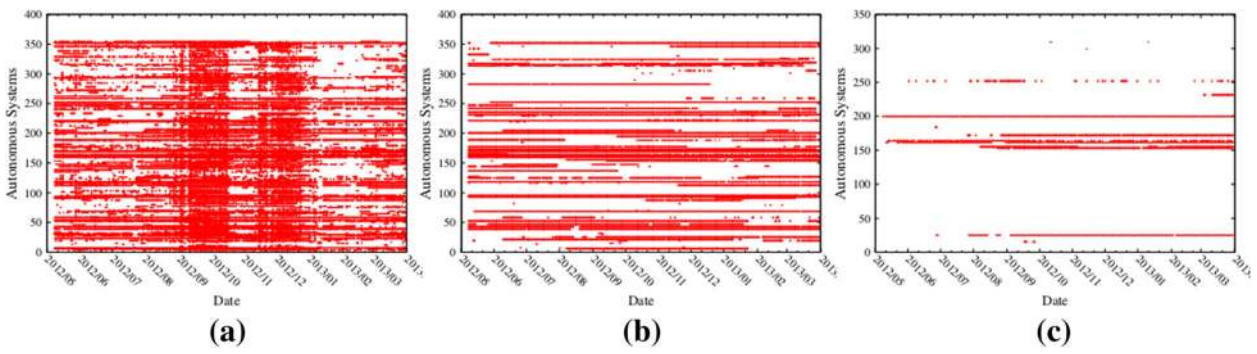

Figure 10 Activity of Autonomous Systems in group 4, per protocol. **(a)** SMTP. **(b)** SOCKS. **(c)** HTTP. ASes appear ordered in the y axis. There is a mark at position (x,y) if AS y was active on date x.

that, in general, consist of a larger number of machines. Observing Figure 12(b), we see an inversion: HTTP or SOCKS *SpamBands* tend to send more messages than SMTP ones. This happens precisely due to the fact previously mentioned. Since those who use SOCKS or HTTP are probably dedicated servers, they use all their resources to send a large number of messages. On the other hand, those who use SMTP and are part of botnets can only send spam moderately, to avoid their identification [20].

6.1 Relationship between *SpamBands* and ASes

In Figure 13, we can observe that the majority of *SpamBands*, for all protocols, have IP addresses from just a few ASes. This result indicates a topological relationship among IP addresses of a *SpamBand*, where machines from the same AS tend to send unwanted messages from the same set of spam campaigns. Furthermore, the small number of *SpamBands* that encompass IP addresses from more than 60 neighborhoods use the SMTP protocol. This result is expected, since SMTP is used by botnets, which tend to have infected machines spread over more ASes.

6.1.1 *SpamBands activities in different neighborhoods*

As already mentioned, Autonomous Systems were classified in four groups, where groups 2 and 4 were considered bad neighborhoods. The results in the chart of Figure 14 shows that half of the *SpamBands* (about 50%) have IP addresses from Autonomous Systems that belong only to groups 2 or 4. Then come SpamBands that use both hosts in neighborhoods of type 2 and 4, and then those that use only group 1. This confirms that the ASes were classified correctly into those four groups and suggests that most of the IP addresses in *SpamBands* are in bad neighborhoods. Thus, this result points out that the efforts against the spam abuse have to focus in Autonomous Systems that are considered as bad neighborhoods.

Furthermore, most of *SpamBands* that contain AS from group 2 usually use the HTTP or SOCKS protocol. This

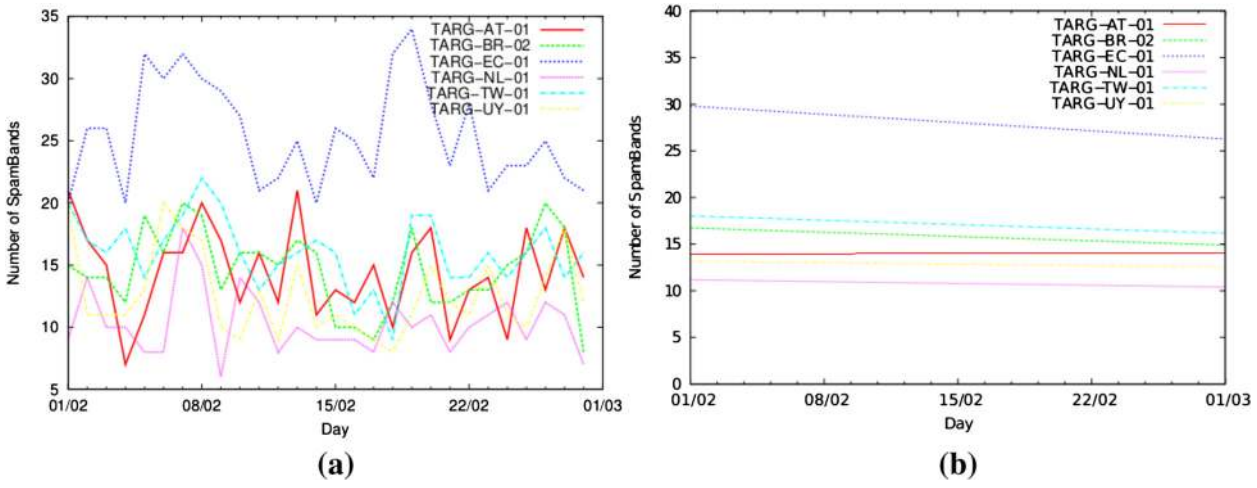

Figure 11 SpamBands active in various honeypots over time. **(a)** The number of SpamBands found each day in eight different honeypots. **(b)** Their linear regressions.

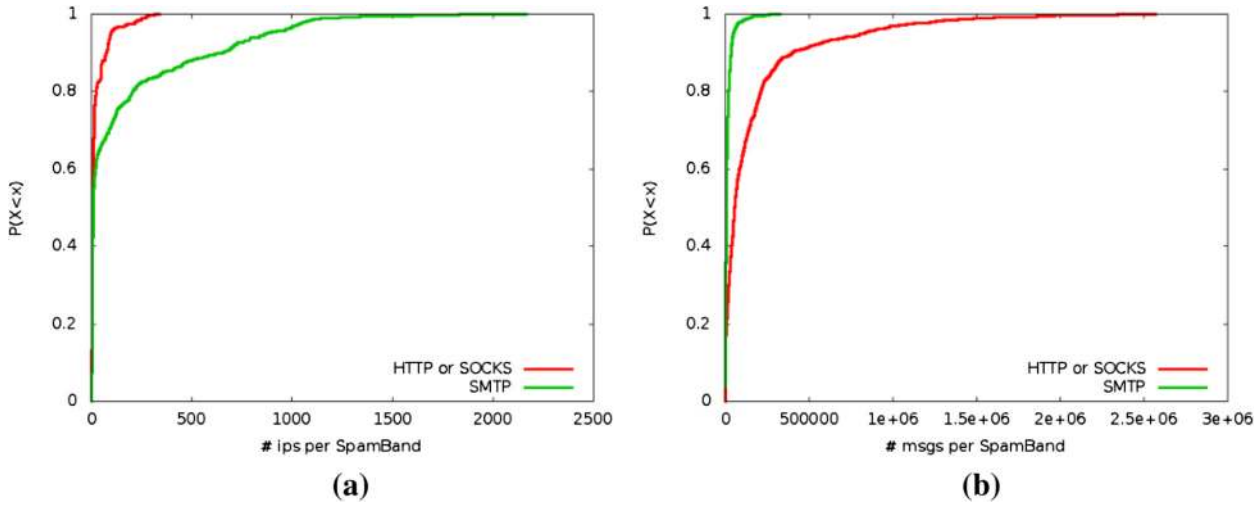

Figure 12 Distribution of IPs present and messages sent by SpamBands. **(a)** CDF of the number of IP addresses per SpamBand. **(b)** CDF of the number of messages sent per SpamBand.

was expected, once group 2 seems to have a lot of dedicated server machines to send spam. On the other hand, the *SpamBands* with AS in group 4 use the SMTP protocol. This result was also expected because most machines in these neighborhoods seems to belong to botnets.

6.1.2 SpamBands clustering

In this section we analyze the clustering coefficient inside the *SpamBands* to verify if IP addresses SpamBands interact more with other IP addresses in their AS than with IP addresses from others neighborhoods. The internal clustering coefficient (ICC) of a SpamBand is the average of the clustering coefficient in each AS considering only the internal connections, *i.e.*, connections among IP addresses that belong to a same Autonomous System. On the other hand, the external clustering coefficient (ECC) of a SpamBand takes the average of the clustering coefficient in each AS considering only the external connections, *i.e.*, connections among IP addresses of different Autonomous Systems.

As shown in Figure 15, 55% of the *SpamBands* have ICC higher than 0.3 while and only 42% have ECC higher than this value. Moreover, as we can see, the standard behavior is that each *SpamBand* have an ECC smaller than its ICC. We conclude that inside a *SpamBand*, the relationships between IP addresses which belong to a same AS are

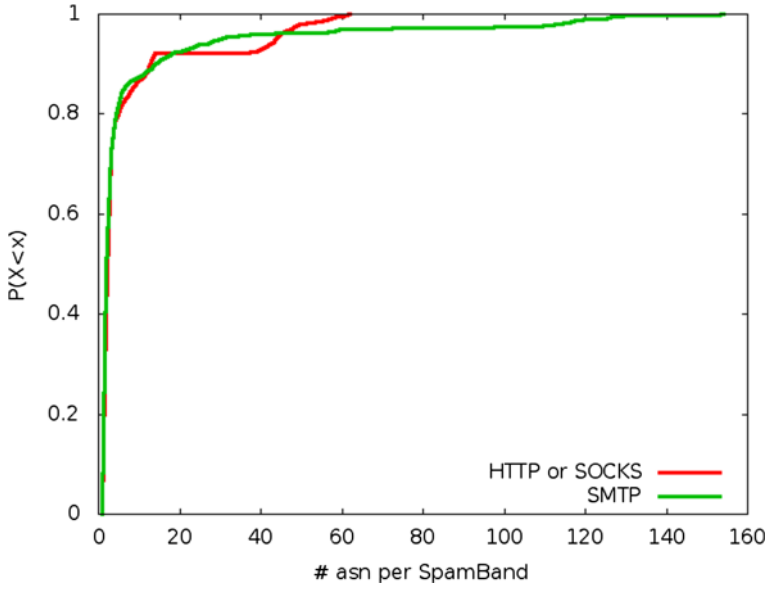

Figure 13 Distribution of the number of AS per SpamBand. Show the distribution of the number of different AS that appears in a same SpamBand.

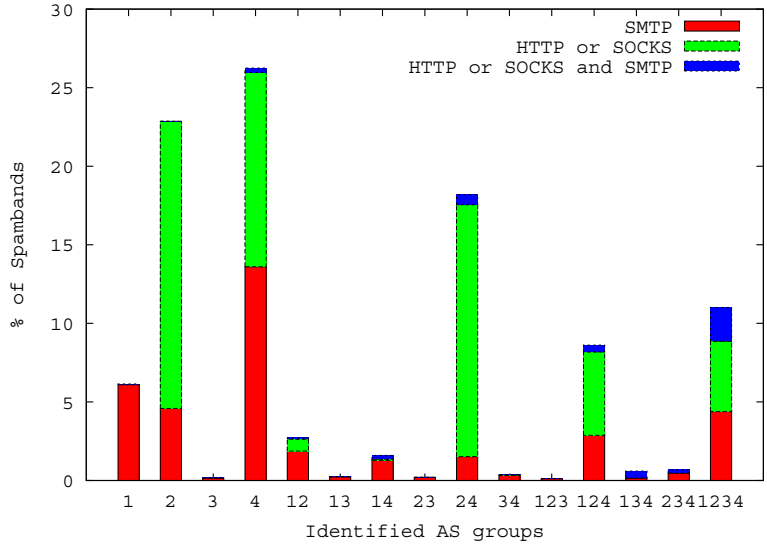

Figure 14 SpamBands distribution over neighborhood types (AS groups). Shows how SpamBands are distributed over neighborhood types (AS groups). The numbers on the x axis identify individual groups (1, 2, 3, 4) or combinations of two or three groups (*e.g.*, 24 means a SpamBand has members in ASes in neighborhood types identified by groups 2 and 4).

more intense. It suggests that there is a topological correlation among IP addresses inside a *SpamBand*, showing that Autonomous Systems are a good way to represent neighborhoods.

7 Conclusions

Several efforts are under way to combat spam, but this task has been made difficult due to the technical sophistication of spammers. This paper tries to shed some light on the sources of spam messages, to help the development of techniques and policies to fight spam at its origin. Our results show that, although spam messages are being sent from various networks, most of the traffic is concentrated in a few Autonomous Systems, and that can be used to identify spam sources and fight them. Moreover, we grouped ASes into four categories based on their spam dissemination behavior. Those groups shown that we can identify good and bad neighborhoods, some with

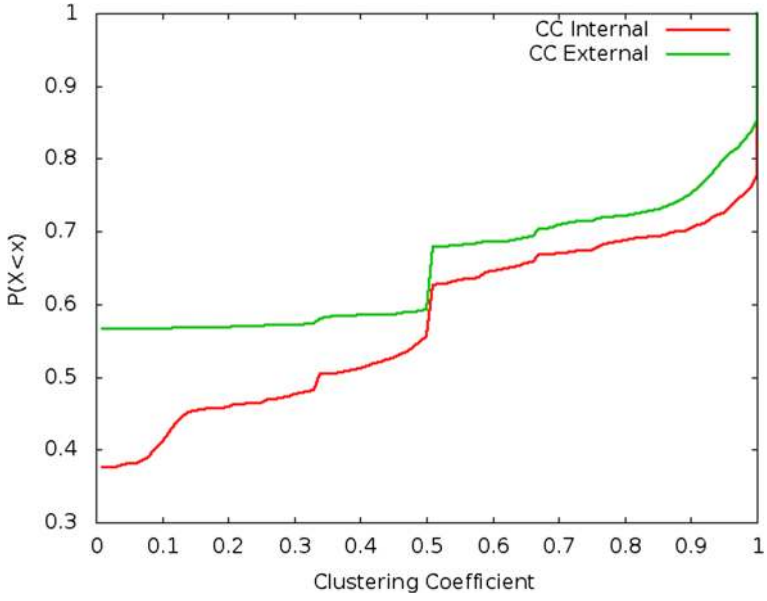

Figure 15 SpamBands clustering within ASes. Verification of the clustering coefficient SpamBand graphs when considered the AS borders: average internal (within AS) and external (among ASes) clustering coefficientes.

many infected machines, others with just a few on-and-off senders that get shut down quickly, other which are conniving with a few heavy spammers.

By identifying machines that participate together in a spam campaign (SpamBands), we observed that most campaigns originated from neighborhoods of a single type, or may include hosts in the two types of heavy sending neighborhoods at the same time. All that can be used to identify major sources of spam to help stop that kind of traffic.

As future work, we plan to conduct further analysis on each of the neighborhood categories found to better understand the differences among them. We also intend to better understand the behavior of the category considered good neighborhoods and check whether security policies used to define the behavior of those autonomous systems can serve as a model to others.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
OF carried out the Neighborhood analysis, participated in the SpamBand analysis, and drafted part of the manuscript. PHBLC participated in the SpamBand and Neighborhood analysis, and drafted part of the manuscript. EF carried out the SpamBand analysis and helped draft the manuscript. DG and WM participated in the design of the study and coordinated it, guided the analysis and helped draft the manuscript. CH, KSJ and MHPC built the collection infrastructure, provided the dataset and participated in the analysis. All authors read and approved the final manuscript.

Acknowledgments
This work was partially funded by NIC.Br, Fapemig, CAPES, CNPq and InWeb.

Author details
[1]Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [2]Brazilian Emergency Response Team (CERT.br), Brazilian Network Information Center (NIC.br), São Paulo, Brazil.

Received: 17 September 2014 Accepted: 29 March 2015
Published online: 11 May 2015

References

1. Orman H (2013) The compleat story of phish. Int Comput IEEE 17(1):87–91
2. Newman MEJ, Forrest S, Balthrop J (2002) Email networks and the spread of computer viruses. Phys Rev E 66:035101
3. Sipior JC, Ward BT, Bonner PG (2004) Should spam be on the menu? Commun ACM 47(6):59–63
4. Guerra PHC, Guedes D, Wagner Meira J, Hoepers C, Chaves MHPC, Steding-Jessen K (2010) Exploring the spam arms race to characterize spam evolution. In: Proceedings of the 7th Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS), Redmond, WA
5. van Wanrooij W, Pras A (2010) Filtering spam from bad neighborhoods. Int J Netw Manag 20(6):433–444
6. Moreira Moura GC, Sadre R, Pras A (2011) Internet bad neighborhoods: the spam case. In: Festor O, Lupu E (eds). 7th International Conference on Network and Services Management (CNSM 2011), Paris, France. IEEE Communications Society, USA. pp 1–8
7. Ramachandran A, Feamster N (2006) Understanding the Network-Level Behavior of Spammers. SIGCOMM Comput Commun Rev 36(4):291–302
8. Duan Z, Gopalan K, Yuan X (2011) An empirical study of behavioral characteristics of spammers: Findings and implications. Comput Commun 34(14):1764–1776
9. Pathak A, Hu YC, Mao ZM (2008) Peeking into spammer behavior from a unique vantage point. In: Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats. LEET'08. USENIX Association, Berkeley, CA, USA. pp 3–139. http://dl.acm.org/citation.cfm?id=1387709.1387712
10. Gomes LH, Almeida RB, Bettencourt LMA, Almeida V, Almeida JM (2005) Comparative Graph Theoretical Characterization of Networks of Spam and Legitimate Email. In: Proceedings of the Second Conference on Email and Anti-Spam - CEAS 2005. CEAS, Stanford, CA, USA
11. Kokkodis M, Faloutsos M (2009) Spamming botnets: Are we losing the war? In: Proceedings of the 6th Conference on E-mail and Anti-spam (CEAS), Mountain View, CA
12. Guerra PHC, Pires DEV, Guedes D, Wagner Meira J, Hoepers C, Steding-Jessen K (2008) A campaign-based characterization of spamming strategies. In: Proceedings of the 5th Conference on E-mail and Anti-spam (CEAS), Mountain View, CA
13. Xie Y, Yu F, Achan K, Panigrahy R, Hulten G, Osipkov I (2008) Spamming botnets: signatures and characteristics. In: Bahl V, Wetherall D, Savage S, Stoica I (eds). SIGCOMM. ACM, Seatle, WA. pp 171–182
14. Fonseca O, Las-Casas PHB, Fazzion E, Guedes D, Jr. WM, Hoepers C, Steding-Jessen K, Chaves MHP (2014) Vizinhanças ou condomínios: uma análise da origem de spams com base na organização de sistemas autônomos. In: Brazilian Symposium on Computer Networks and Distributed Systems (SBRC) (In Portuguese), Florianópolis, Brazil
15. Fazzion E, Las-Casas PHB, Fonseca O, Guedes D, Jr. WM, Hoepers C, Steding-Jessen K, Chaves MHP (2014) Spambands: Uma metodologia para identificação de fontes de spam agindo sob uma coordenação. In: Brazilian Symposium on Information Security and Computer Systems (SBSeg) (In Portuguese), Belo Horizonte, Brazil
16. Steding-jessen K, Vijaykumar NL, Montes A (2008) Using Low-Interaction Honeypots to Study the Abuse of Open Proxies to Send Spam. INFOCOMP J Comput Sci 7(1):45–53
17. Totti LC, Moreira REA, Fazzion E, Fonseca O, Wagner Meira J, Guedes D, Hoepers C, Steding-Jessen K, Chaves MHP (2012) Impacto da Evolução Temporal na Detecção de Spammers na Rede de Origem. In: SBRC 2012, Ouro Preto, Brasil
18. Almeida H, Guedes D, Meira W, Zaki MJ (2011) Is there a best quality metric for graph clusters? In: Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I, Athens, Greece. pp 44–59
19. Pelleg D, Moore AW (2000) X-means: Extending k-means with efficient estimation of the number of clusters. In: ICML, San Francisco, CA. pp 727–734
20. John JP, Moshchuk A, Gribble SD, Krishnamurthy A (2009) Studying Spamming Botnets Using Botlab. In: 6th USENIX Symp. on Networked Systems Design and Implementation, Boston, EUA