Characterizing a Spam Traffic*

Luiz Henrique Gomes, Cristiano Cazita
Jussara M. Almeida, Virgílio Almeida, Wagner Meira Jr.

Department of Computer Science
Federal University of Minas Gerais
Belo Horizonte - Brazil
{lhg, cazita, jussara, virgilio, meira}@dcc.ufmg.br

ABSTRACT

The rapid increase in the volume of unsolicited commercial e-mails, also known as spam, is beginning to take its toll in system administrators, business corporations and end-users. Widely varying estimates of the cost associated with spam are available in the literature. However, a quantitative analysis of the determinant characteristics of spam traffic is still an open problem. This work fills this gap and presents what we believe to be the first extensive characterization of a spam traffic.

As basis for our characterization, standard spam detection techniques are used to classify over 360 thousand incoming e-mails to a large university into two categories, namely spam and non-spam. For each of the two resulting workloads, as well as for the aggregate workload, we analyze a set of parameters, aiming at identifying the characteristics that significantly distinguish spam from non-spam traffic, assessing the qualitative impact of spam on the aggregate traffic and, possibly, drawing insights into the design of more effective spam detection techniques.

Our characterization reveals significant differences in the spam and non-spam traffic patterns. E-mail arrival process, size distribution as well as the distributions of popularity and temporal locality of e-mail recipients are key workload aspects which distinguish spam from traditional e-mail traffic. We conjecture that these differences are consequence of the inherently different mode of operation of spam and non-spam senders. Whereas non-spam e-mail transmissions are typically driven by social bilateral relationships, spam transmission is usually a unilateral action, based solely on the senders's will to reach as many users as possible.

Categories and Subject Descriptors

[Computer Systems Organization (Performance of Systems)]: Measurement techniques, Modeling techniques.

General Terms

Measurement, Performance

Keywords

Workload Characterization, SPAM, E-mail Traffic

1. INTRODUCTION

E-mail has become a *de facto* means to disseminate information to millions of users in the Internet. However, the volume of unsolicited e-mails containing, typically, commercial content, also known as *spam*, is increasing at a very fast rate. In September 2001, 8% of all e-mails in US were spams. By July 2002, this fraction had increased to 35% [1]. More recent studies report that, in North America, a business user received on average 10 spams per day in 2003, and that this number is expected to grow by a factor of four by 2008 [2]. Furthermore, AOL and MSN, two large ISPs, report blocking, daily, a total of 2.4 billion spams from reaching their customers' in boxes. This traffic corresponds to about 80% of daily incoming e-mails at AOL [3].

This rapid increase in spam traffic is beginning to take its toll in end users, business corporations and system administrators. Results from a recent survey with over six thousand American e-mail users report that over 50% of them are less trusting of e-mail systems, and over 70% of them believe being online has become unpleasant or annoying due to spam [3]. The impact of spam traffic on the productivity of workers of large corporations is also alarming. Research firms estimate the yearly cost per worker at anywhere from US$ 50 to US$ 1400, and the total annual cost associated with spam to American businesses in the range of US$ 10 billion to US$ 87 billion [3]. Finally, estimates of the cost of spam must also take into account the costs of computing and network infra-structure upgrades as well as quantitative measures of its impact on the quality of service available to traditional non-spam e-mail traffic and other "legitimate" Internet applications.

A number of approaches have been proposed to alleviate the impact of spam. These approaches can be categorized into pre-acceptance and post-acceptance methods, based on whether they detect and block spam before or after accepting the e-mail [4]. Examples of pre-acceptance methods are black lists [5] and gray lists or tempfailing [6]. Pre-acceptance approaches based on server authentication [7, 8] and accountability [9] have also been recently proposed. Examples of post-acceptance methods include bayesian filters [10], collaborative filtering [11] and e-mail prioritization [4].

Although existing spam detection and filtering techniques have, reportedly, very high success rates (up to 97% of spams are detected

*Jussara M. Almeida, Virgílio Almeida and Wagner Meira Jr. are supported by CNPq/Brazil. Luiz H. Gomes is supported by Central Bank of Brazil

IMC'04, October 25–27, 2004, Taormina, Sicily, Italy.

[11]), they suffer from two limitations. First, the rate of false positives, i.e., legitimate e-mails classified as spams, can be as high as 15% [12], incurring costs that are hard to mensurate. Second, the lifetime of existing techniques is compromised by spammers frequently changing their mode of operation (e.g., forging their e-mail addresses and/or misspelling in spam messages). In other words, spam filters have their effectiveness frequently challenged. Constant upgrades and new developments are necessary.

Despite the large number of reports on spam cost and the plethora of previously proposed spam detection and filtering methods, a quantitative analysis of the determinant characteristics of this type of Internet traffic is still in demand. In addition to some previous e-mail workload characterizations [13, 14], we are aware of only two limited efforts towards analyzing some characteristics of spam traffic in the literature [4, 7].

This paper takes an innovative approach towards addressing the problems caused by spam and presents what we believe to be the first extensive characterization of a spam traffic. Our goal is to develop a deep understanding of the fundamental characteristics of spam traffic and spammer's behavior, in hope that such knowledge can be used, in the future, in the design of more effective techniques for detecting and combating spams.

Our characterization is based on an eight-day log of over 360 thousand incoming e-mails to a large university in Brazil. Standard spam detection techniques are used to classify the e-mails into two categories, namely, spam and non-spam. For each of the two resulting workloads, as well as for the aggregate workload, we analyze a set of parameters, based on the information available in the e-mail headers. We aim at identifying the quantitative and qualitative characteristics that significantly distinguish spam from non-spam traffic and assessing the impact of spam on the aggregate traffic by evaluating how the latter deviates from the non-spam traffic.

Our key findings are:

- Unlike traditional non spam e-mail traffic, which exhibits clear weekly and daily patterns, with load peaks during the day and on weekdays, the numbers of spam e-mails, spam bytes, distinct active spammers and distinct spam e-mail recipients are roughly insensitive to the period of measurement, remaining mostly stable during the whole day, for all days analyzed.

- Spam and non spam inter-arrival times are exponentially distributed. However, whereas the spam arrival rates remain roughly stable across all periods analyzed. The arrival rates of non spam e-mails vary as much as a factor of five in the periods analyzed.

- E-mail sizes in the spam, non-spam and aggregate workloads follow Lognormal distributions. However, in our workload the average size of a non-spam e-mail is from six to eight times larger than the average size of a spam. Moreover, the coefficient of variation (CV) of the sizes of non-spam e-mails is around three times higher than the CV of spam sizes. The impact of spam on the aggregate traffic is a decrease on the average e-mail size but an increase in the size variability.

- The distribution of the number of recipients per e-mail is more heavy-tailed in the spam workload. Whereas only 5% of non-spam e-mails are addressed to more than one user, 15% of spams have more than one recipient, in our workload. In the aggregate workload, the distribution is heavily influenced by the spam traffic, deviating significantly from the one observed in the non-spam workload.

- Regarding daily popularity of e-mail senders and recipients, the main distinction between spam and non-spam e-mail traffics comes up in the distribution of the number of e-mails per recipient. Whereas in the non-spam and aggregate workloads, this distribution is well modeled by a single Zipf-like distribution plus a constant probability of a user receiving only one e-mail per day, the distribution of the number of spams a user receives per day is more accurately approximated by the concatenation of two Zipf-like distributions, in addition to the constant *single-message* probability.

- There are two distinct and non-negligible sets of non-spam recipients: those with very strong temporal locality and those who receive e-mails only sporadically. These two sets are not clearly defined in the spam workload. In fact, temporal locality is, on average, much weaker among spam recipients and even weaker among recipients in the aggregate workload. Similar trends are observed for the temporal locality among e-mail senders.

Therefore, our characterization reveals significant differences between the spam and non-spam workloads. These differences are possibly due to the inherent distinct nature of e-mail senders and their connections with e-mail recipients in each group. Whereas a non-spam e-mail transmission is the result of a bilateral relationship, typically initiated by a human being, driven by some social relationship, a spam transmission in basically a unilateral action, typically performed by automatic tools and driven by the spammer's will to reach as many targets as possible, indiscriminately, without being detected.

The remaining of this paper is organized as follows. Section 2 discusses related work. Our e-mail workloads and the characterization methodology are described in Section 3. Section 4 analyzes temporal variation patterns in the workloads. E-mail traffic characteristics are discussed in Section 5. E-mail recipients and senders are analyzed in Section 6. Finally, Section 7 presents conclusions and directions for future work.

2. RELATED WORK

Developing a clear understanding of the workload is a key step towards the design of efficient and effective distributed systems and applications. A number of characterizations and analyses of different workload types which led to valuable insights into system design are available in the literature, including the characterization of web workloads [15], streaming media workloads [16, 17, 18] and, recently, peer-to-peer [19] and chat room workloads [20]. To the best of our knowledge, no previous work has performed a thorough characterization of spam traffic. Next we discuss previous characterizations and analyses of e-mail workload [4, 7, 13, 14].

In [13, 14], the authors provide an extensive characterization of several e-mail server workloads, analyzing e-mail inter-arrival times, e-mail sizes, and number of recipients per e-mail. They also analyze user accesses to mail servers (through the POP3 protocol), characterizing inter-access times, number of messages per user mailboxes, mailbox sizes and size of deleted e-mails, and propose models of user behavior. In this paper, we characterize not only a general e-mail workload, but also a spam workload, in particular, aiming at identifying a signature of spam traffic, which can be used in the future for developing more effective spam-controlling techniques. In Sections 4-5, we contrast our characterization results for non-spam e-mails with those reported in [13, 14].

Twining *et al* [4] present a simpler server workload characterization, as the starting point for investigating the effectiveness of novel techniques for detecting and controlling junk e-mails (i.e., virus and spams). They analyze the logs of two e-mail servers that include a virus checker and a spam filter, and characterize the arrival process of each type of e-mail (spam, virus, and “good”) per-sender, the percentage of servers that send only junk e-mails, only good e-mails, and a mixture of both. A major conclusion of the paper is that popular spam detection mechanisms such as blacklist, tempfailing, and rate-limiting are rather limited in handling the problem. This paper presented a more thorough characterization of spam traffic, and contrasts, whenever appropriate, our findings to those found in [4] .

In [7], the authors analyze the temporal distribution of spam arrivals as well as spam content at selected sites from the AT&T and Lucent backbones. They also discuss the factors that make users and domains more likely to receive spams and the reasons that lead to the use of spam as a communication and marketing strategy. The paper also includes a brief discussion of the pros and cons of several anti-spam strategies.

3. E-MAIL WORKLOAD

This section introduces the e-mail workload analyzed in this paper. Section 3.1 describes the data source and collection architecture. The methodology used in the characterization process is presented in Section 3.2. Section 3.3 provides an overview of our e-mail workload.

3.1 Data Source

Our e-mail workload consists of anonymized SMTP logs of incoming e-mails to a large university, with around 22 thousand students, in Brazil. The logs are collected at the central Internet facing e-mail server of the university. This server handles all e-mails coming from the outside and addressed to most students, faculty and staff, with e-mail addresses under the major university domain name. Only the e-mails addressed to two out of over 100 university subdomains (i.e., departments, research labs, research groups) do not pass through and, thus, are not logged by, the central server.

The central e-mail server runs the Exim e-mail software [21], the Amavis virus scanner [22] and the Trendmicro Vscan anti-virus tool [23]. It also runs a set of pre-acceptance spam filters, including local black lists and local heuristics for detecting suspicious senders. These filters block on average 50% of all daily SMTP connection arrivals. The server also runs SpamAssassin [24], a popular spam filtering software, over all e-mails that are accepted. SpamAssassin detects and filters spams based on a set of user-defined rules. These rules assign scores to each received e-mail based on the presence in the subject or in the e-mail body of one or more pre-categorized keywords taken from a constantly changing list. High ranked e-mails are flagged as spams. SpamAssassin also uses size-based rules, which categorize messages larger than a pre-defined size as legitimate non-spam e-mails. E-mails that are neither flagged as spam nor as virus-infected are forwarded to the appropriate local servers, indicated by the sub-domain names of the recipient users.

We analyze an eight-day log collected by the Amavis software at the central e-mail server, during academic year at the university. Our logs store the header of each e-mail that passes the pre-acceptance filters, along with the results of the tests performed by SpamAssassin and the virus scanners. In other words, for each e-mail that is accepted by the server, the log contains the arrival time, the size, the sender e-mail address, a list of recipient e-mail addresses and flags indicating whether the e-mail was classified as spam and whether it was detected to be infected with a virus. Figure 1 shows the overall data collection architecture at the central e-mail server.

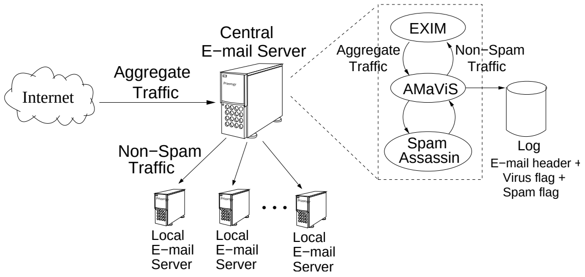


Figure 1: Data Collection the Central E-mail Server

E-mails that are flagged with virus or addressed to recipients in a domain name outside the university, for which the central e-mail server is a published relay, are *not* included our analysis. These e-mails correspond to only 0.8% of all logged data.

Note that the central server does not perform any test on the existence of the recipient addresses of the accepted e-mails. Such tests are performed by the local servers. Thus, some of the recipient e-mail addresses in our logs may not actually exist. These recipient addresses could be result of honest mistakes or the consequence of dictionary attacks [25], a technique used by some spammers to automatically generate a target distribution list with a large number of *potential* e-mail addresses.

3.2 Characterization Methodology

As basis for our characterization, we first group the e-mails logged by Amavis into two categories, namely, *spam* and *non-spam* (also referred to as “ham” in the literature [26]), based on whether the e-mail was flagged by SpamAssassin. Three distinct workloads are then defined:

- *Spam* - only e-mails flagged by SpamAssassin.
- *Non-Spam* - only e-mails not flagged by SpamAssassin.
- *Aggregate* - all e-mails logged by Amavis.

We characterize each workload separately. The purpose is threefold. First, we can compare and validate our findings for the non-spam workload with those reported in previous analyzes of traditional (non-spam) e-mail traffic [4, 13, 14, 27]. Second, we are able to identify the characteristics that significantly distinguish spam from non-spam traffic. Finally, we are also able to assess the quantitative and qualitative impact of spam on the overall e-mail traffic, by evaluating how the aggregate workload deviates from the non-spam workload.

Our characterization focuses on the information available in the e-mail headers, logged by Amavis. In other words, we characterize the e-mail arrival process, distribution of e-mail sizes, distribution of the number of recipients per e-mail, popularity and temporal locality among e-mail recipients and senders. Characterization of e-mail content is left for future work.

Each workload aspect is analyzed separately for each day in our eight-day log, recognizing that their statistical characteristics may vary with time. The e-mail arrival process is analyzed during periods of approximately stable arrival rate, as daily load variations may also impact the aggregate distribution.

To find the distribution that best models each workload aspect, on each period analyzed, we compare the least square differences of the best fitted curves for a set of alternative distributions commonly found in other characterization studies [13, 16, 17, 18, 20, 28, 29, 30]. We also visually compared the curve fittings at the body and

Table 1: Summary of the Workloads (CV = Coefficient of Variation)

Measure	Non-Spam	Spam	Aggregate
Period	2004/01/19-26	2004/01/19-26	2004/01/19-26
Number of days	8	8	8
Total # of e-mails	191,417	173,584	365,001
Total size of e-mails	11.3 GB	1.2 GB	12.5 GB
Total # of distinct senders	12,338	19,567	27,734
Total # of distinct recipients	22,762	27,926	38,875
Avg # distinct recipients/msg (CV)	1.1 (0.74)	1.7 (1.38)	1.4 (1.27)
Avg # msgs/day (CV)	23,927 (0.26)	21,698 (0.08)	45,625 (0.17)
Avg # bytes/day (CV)	1.5 GB (0.39)	164 MB (0.19)	1.7 GB (0.37)
Avg # distinct senders/day (CV)	3,190 (0.22)	5,884 (0.10)	8,411 (0.11)
Avg # distinct recipients/day (CV)	8,981 (0.15)	14,936 (0.24)	19,935 (0.20)

at the tail of the measured data, favoring a better fit at either region whenever appropriate to capture the most relevant aspects of the workload to system design. For instance, shorter inter-arrival times and larger e-mail sizes have a stronger impact on server capacity planning. Thus, we favor a better fit at the body (tail) of the data for determining the arrival process (distribution of e-mail sizes). In Sections 5-6, we show only the results for the best fits.

By visually inspecting the list of sender *user names* in our spam workload, we found that a large number of them seemed a random sequence of characters, suggesting forging. Note that sender IP addresses may also be forged, although we expect it to happen less frequently. Our logs contain only sender domain names. However, sender IP addresses are separately collected by the Exim software. By analyzing the Exim logs collected at the same period of our Amavis logs, we found that, on average, a single sender domain name is associated with 15 different IP addresses, whereas the average number of different domains per sender IP address is only 6. In other words, there is no indication of which information is more reliable. Because the results of SpamAssassin are available only in the Amavis logs and a merge of both logs is hard to build, our per-sender analysis focuses only on sender domain names. Thus, throughout this paper, we simply use:

- *E-mail sender* - to refer to the e-mail sender domain.
- *E-mail recipient* - to refer to an e-mail recipient user name.

Table 2: Distribution of Senders and Recipients

	Senders		Recipients	
Group	%	% Msg	%	% Msg
Only Non-Spam	29	31	24	10
Only Spam	56	23	38	20
Mixture	15	46	37	70

3.3 Overview of the Workloads

An overview of our three workloads is provided in Table 1. Note that although spams correspond to almost 50% of all e-mails, spam traffic corresponds to only 10% of all bytes received during the analyzed period. Furthermore, the total number of distinct spammers is almost 60% larger than the number of distinct senders in the non-spam workload. Thus, the average number of e-mails originating from the same domain is smaller in the spam workload, possibly due to spammers periodically changing their e-mail domain names to escape from black lists. Note that the total number of spam recipients as well as the number of recipients per spam are also significantly larger than the corresponding metrics in the non-spam workload. This may be explained by spammers' will to target as many address as possible (e.g., dictionary attacks). Another interesting point is the much lower variability in spam traffic, which is further discussed in Section 4. Similar conclusions hold on a daily basis, as shown in the last 5 rows in Table 1.

Table 2 shows the percentages of senders and recipients that send and receive only non-spam e-mails, only spams and a mixture of both. It also shows the percentage of e-mails each category of sender/recipient is responsible for. More than half of all domains send only spams whereas 15% of them send both types of e-mails. We also point out that, on average, six out of the ten most active spam senders on each day send only spams. Nevertheless, the spam-only servers are responsible for only 23% of all e-mails, whereas 46% of the e-mails originate from domains that send a mixture of spams and non-spams. These results may be explained by spammers frequently "forging" new domains. In [4], the authors also found a large fraction of senders who send only junk (virus, spam) e-mails. However, they found those senders accounted for a larger fraction of the e-mails in their workloads.

Table 2 also shows that whereas 24% of all recipients are not target of spam, around 38% of them appear only in the spam workload and receive 20% of all e-mails, in our log. Furthermore, we found that around 50% of the spam-only recipients received less than 5 e-mails during the whole log, and that a number of them seemed forged (e.g., randomly generated sequence of characters). These observations lead us to speculate that many spam-only recipients are the result of two frequent spammer actions: dictionary attacks and removal of recipients from their target list after finding they do not exist (i.e., after receiving a "not a user name" SMTP answer). They also illustrate a potentially harmful side-effect of spam, which is the use of network and computing resources for transmitting and processing e-mails that are addressed to non-existent users and, thus, that will be discarded only once they reach the local e-mail server they are addressed to.

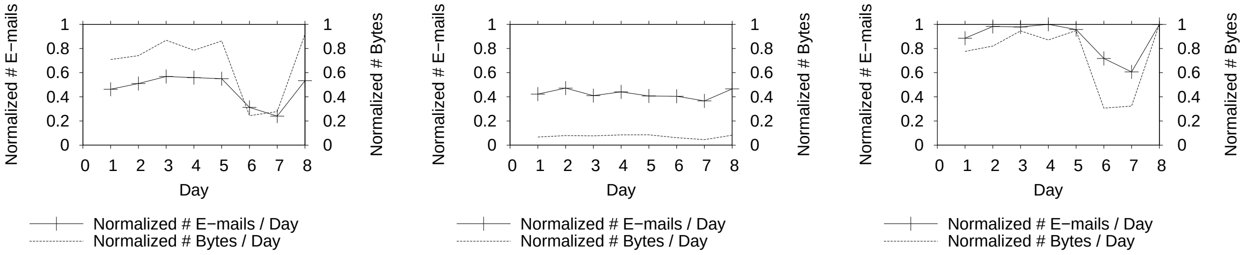
Interestingly, we found that all spams in our workload are addressed to existing domain names under the university major domain name. Note that there exists over a hundred different subdomain names under the university major domain name. Thus, spammers seem to perform dictionary attacks by starting from a real domain name and guessing a list of possible user names in order to maximize the chance of hitting a real user. An evaluation of the correlation between the spam content and the target domain name to assess the knowledge of spammers with respect to their targets is left for future work.

4. TEMPORAL VARIATION PATTERNS IN E-MAIL TRAFFIC

This section discusses temporal variation patterns in each of our three e-mail workloads, namely spam, non-spam and aggregate workloads. Section 4.1 analyzes daily and hourly variations in load

Table 3: Summary of Hourly Load Variation

Workload	Metric	Minimum	Maximum	Average	CV
Non-Spam	# E-mails/Hour	232 - 435	703 - 4,676	513 - 1,213	0.20 - 0.74
	# Bytes/Hour (MB)	4 - 11	46 - 349	23 - 86	0.45 - 0.98
Spam	# E-mails/Hour	194 - 776	1,081 - 2,086	781 - 1,007	0.12 - 0.36
	# Bytes/Hour (MB)	1.7 - 5.7	6.1 - 18.4	4.3 - 8.0	0.15 - 0.45
Aggregate	# E-mails/Hour	500 - 1,210	1,681 - 6,762	1,294 - 2,134	0.13 - 0.55
	# Bytes/Hour (MB)	8.7 - 16.8	50 - 367	29 - 93	0.36 - 0.93

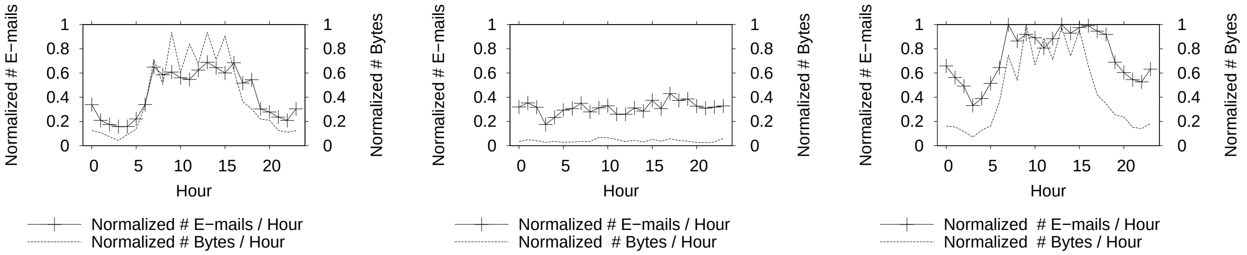


(a) Non-Spam

(b) Spam

(c) Aggregate

Figure 2: Daily Load Variation (Normalization Parameters: Max # E-mails = 51,226, Max # Bytes 2.24 GB)



(a) Non-Spam

(b) Spam

(c) Aggregate

Figure 3: Hourly Load Variation (Normalization Parameters: Max # E-mails = 2,768, Max # Bytes 197 MB)

intensity, measured in terms of the total number of e-mails and total number of bytes. Temporal variations in the numbers of distinct e-mail recipients and senders are discussed in Section 4.2.

4.1 Load Intensity

Figure 2 shows daily load variations in the number of messages and number of bytes, for non-spam, spam and agregate workload respectively. The graphics show load measures normalized by the peak daily load observed in the aggregate traffic. The normalization parameters are shown in the caption of the figure.

Figure 2-a shows that the daily load variations in the non-spam e-mail traffic exhibit the traditional bell-shape behavior, typically observed in other web workloads [16, 18, 17], with load peaks during weekdays and a noticeable decrease in load intensity over the weekend (days six and seven). On the other hand, Figure 2-b shows that spam traffic does not present any significant daily variation. The daily numbers of e-mails and bytes are roughly uniformly distributed over the whole week. This stability in the daily spam traffic was previously observed in [7] for a much lighter workload, including only 5% of all e-mails received. Figure 2-c shows that the impact of this distinct behavior on the aggregate traffic is a smoother variation in the number of e-mails per day. The variation in the aggregate number of bytes, on the other hand, has a pattern very similar to the one observed in the non-spam workload, as shown in Figure 2-c. This is because non-spam e-mails account for over 90% of all bytes received (see Table 1).

The same overall behavior is observed for the hourly load variations, as illustrated in Figure 3, for a typical day. Like in [13, 14], traditional non-spam e-mail traffic (Figure 3-a) presents two distinct and roughly stable regions: a high load diurnal period, typically from 7AM to 7PM, (i.e., working hours), during which the central server receives between 65% and 73% of all daily non-spam e-mails, and a low load period covering the evening, night and early morning. On the other hand, the intensity of spam traffic (Figure 3-b) is roughly insensitive to the time of the day: the fraction of spams that arrives during a typical diurnal period is between 50% and 54%. Figure 3-c shows that, as observed for daily load variations, the impact of spam on the aggregate traffic is a less pronounced hourly variation of the number of e-mails received.

Table 3 summarizes the observed hourly load variation statistics. For each workload, it presents the ranges for minimum, maximum, average and coefficient of variation of the number of e-mails and number of bytes received per hour, on each day. Note the higher variability in the number of e-mails and number of bytes in the non-spam workload. Also note that, for any of the three workloads, a higher coefficient of variation is observed in the number of bytes, because of the inherent variability of e-mail sizes. Finally, note that these results are consistent with those of Table 1 for the daily variation in the number of e-mails and bytes in each workload.

Table 4: Summary of Hourly Variation of Number of Recipients and Senders

Workload	Metric	Minimum	Maximum	Average	CV
Non-Spam	# Distinct Recipients/Hour	228 - 383	589 - 2,883	411 - 978	0.21 - 0.58
	# Distinct Senders/Hour	107 - 136	225 - 937	160 - 332	0.14 - 0.61
Spam	# Distinct Recipients/Hour	485 - 1,174	1,397 - 4,095	955 - 2,371	0.15 - 0.41
	# Distinct Senders/Hour	147 - 406	548 - 925	433 - 577	0.10 - 0.24
Aggregate	# Distinct Recipients/Hour	828 - 1,672	2,480 - 6,580	1,505 - 3,179	0.20 - 0.41
	# Distinct Senders/Hour	256 - 541	828 - 1,614	623 - 885	0.12 - 0.33

Qualitative similar results were also found for load variations on a minute basis. The coefficients of variation of the number of e-mails per minute vary in the ranges of 0.45-0.78 and 0.46-0.91 for the spam and non-spam workloads, respectively. The coefficients of variation of the number of bytes per minute are in the ranges of 0.70-1.03 and 1.37-1.75, in the two workloads.

One key conclusion we reach is that, on various time scales, whereas traditional e-mail traffic is concentrated on diurnal periods, the arrival rate of spam e-mails is roughly stable over time. One question that comes up is whether this difference is also observed on a per-sender basis. We analyzed the hourly traffic generated by each of the 50 most active spam-only senders and strictly non-spam senders (see Table 2). We found that each of the 50 strictly spam senders sent, on average, 53% of its daily e-mails during the day. In contrast, the strictly non-spam senders selected concentrate their activity, sending, on average, 63% of their e-mails, at the same period. Similar results were obtained for the 100, 200 and 500 most active senders in each group. Therefore, each spammer independently sends almost half of its e-mails over night, when computing and networking resources are mostly idle. We conjecture that, by using automatic tools, spammers try to maximize their short-term throughput, sending at the fastest rate they can get through without being noticed, throughout the day.

The fundamental difference between spam and non-spam traffic discussed in this section, may be explained by the inherently distinct nature of their sources. Spammers are driven by the goal of reaching as many targets as possible, without being detected. To do so, they use automatic tools to (roughly) uniformly spread the flooding of e-mails over time to avoid being noticed. Thus, a spam transmission is basically a unilateral action. The transmission of a legitimate non-spam e-mail, on the other hand, is the result of a bilateral relationship [28, 29]. It is typically initiated by a human being, driven by some social reason (i.e., work, leisure), during his/her active hours.

4.2 Distinct Senders and Recipients

This section analyzes temporal variations in the numbers of distinct senders and recipients in our workloads. Daily variations and hourly variations for a typical day are shown in Figures 4 and 5, respectively. As before, we show normalized measures, expressed as fractions of the peak aggregate number of senders and recipients in the period. The normalization parameters are given in the captions of the figures.

As observed in the load variation, temporal variations in the number of distinct e-mail senders in the spam workload present significantly different behavior from those observed in the non-spam e-mail workload. Whereas the number of distinct legitimate e-mail senders does present weekly patterns, the number of distinct spammers is roughly stable over the eight days (with a slight increase by the 7^{th} day), as shown in Figures 4-a and 4-b. This difference is even more striking on a hourly basis, as shown in Figures 5-a and 5-b. Again, we speculate that the inherently different nature of the e-mail senders in each workload (automatic tools versus human beings) are responsible for it. In the aggregate traffic, the significant variations observed in the non-spam e-mail traffic are somewhat smoothed out by the roughly stable number of spammers, as shown in Figures 4-c and 5-c.

Regarding the daily variations in the number of distinct recipients, shown in Figure 4, no clear distinction between spam and non-spam traffic was observed. Surprisingly, we found that the number of distinct spam recipients actually decreases significantly by the fourth day. We could not find any reason to explain this weird behavior and plan to look further into that as future work. However, on a hourly basis, we found that whereas the number of distinct recipients of legitimate non-spam e-mails is higher during the day, the number of distinct spam recipients is roughly stable over time, as illustrated in Figure 5, for a typical day. These results are summarized in Table 4, which shows, for each of the three workloads, the observed ranges for the minimum, maximum, average and coefficient of variation of the number of distinct recipients and number of distinct senders per hour.

We also measured the correlation between the number of distinct senders and the number of distinct recipients per hour, on each day, for the three workloads. We found coefficients of correlation between 0.90 and 0.99 in the non-spam workload, and between 0.58 and 0.89 in the spam workload. The lower correlation seems to indicate that there is a larger overlap in the distribution lists of typical spammers. This overlap may be the result of spammers using similar automatic tools to create their targets, trading their distribution lists to extend their reach and/or obtaining the same distribution list from existing web services [31, 32]. As discussed in the previous section, traditional e-mail senders, on the other hand, are driven mostly by social relationships [28, 29]. Thus, shared recipients are most probably due to the fixed number of recipients, who are members of a somewhat closed community (the university).

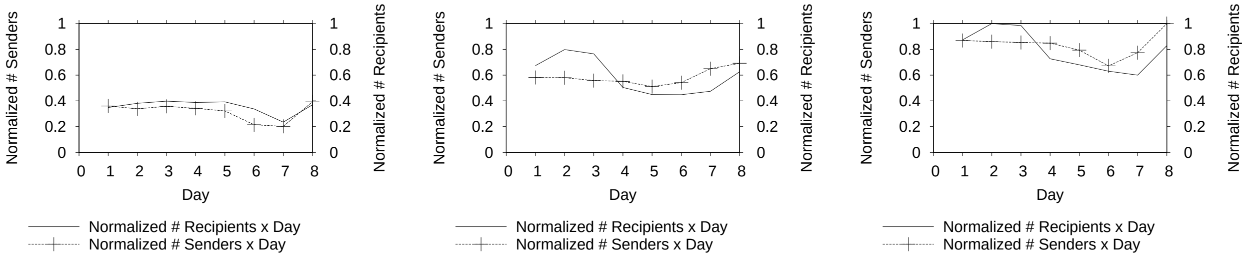
In summary, our results show that, unlike traditional non-spam e-mail traffic, which exhibits clear daily patterns, with load peaks during the day, the numbers of spam e-mails, spam bytes, distinct active spammers and distinct spam recipients are roughly insensitive to the period of measurement, remaining mostly stable during the whole day, for all days analyzed.

5. E-MAIL TRAFFIC CHARACTERISTICS

This section analyzes the characteristics of e-mail traffic for the spam, non-spam and aggregate workloads. The e-mail arrival process is characterized in Section 5.1. The distributions of e-mail sizes and number of recipients per e-mail are analyzed in Sections 5.2 and 5.3, respectively. For each workload characteristics, we discuss the differences between spam and non-spam, pointing out the impact of the former on the aggregate workload.

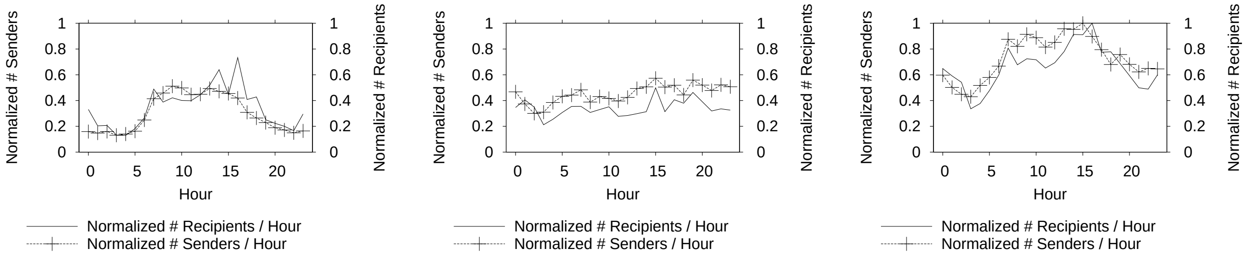
5.1 E-mail Arrival Process

In this section, the e-mail arrival process in each workload is characterized during periods of roughly stable arrival rate, in order to avoid spurious effects due to data aggregation. In the spam workload, such periods are typically whole days, whereas in the



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4: Daily Variation of Number of Senders and Recipients (Normalization Parameters: Max # Senders = 10,089, Max # Recipients = 25,218)



(a) Non-Spam (b) Spam (c) Aggregate

Figure 5: Hourly Variation of Number of Senders and Recipients (Normalization Parameters: Max # Senders = 956, Max # Recipients = 2,802)

non-spam and aggregate workloads, different stable periods are observed during the day and over night.

Table 5: Summary of the Distribution of Inter-Arrival Times

	Inter-Arrival Times		Exponential
Workload	Mean (sec)	CV	Parameter λ
Non-Spam	2.1 - 9.7	1.12 - 1.90	0.10 - 0.48
Spam	3.6 - 4.9	1.08 - 1.99	0.21 - 0.26
Aggregate	1.3 - 3.2	1.07 - 1.73	0.31 - 0.75

Exponential (PDF): $p_X(x) = \lambda e^{-\lambda x}$.

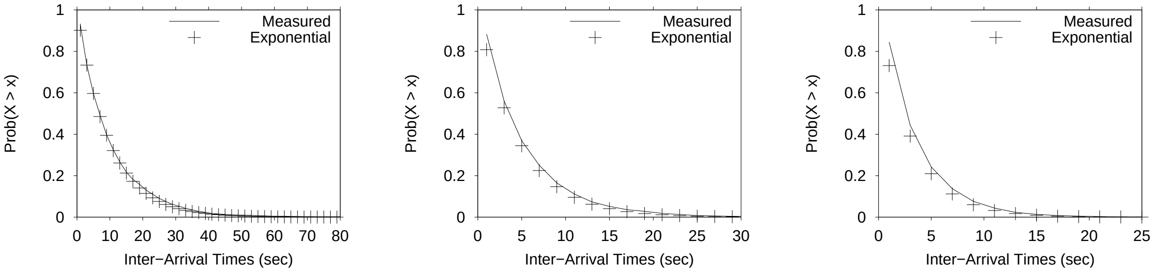
We found that e-mail inter-arrival times are exponentially distributed in all three workloads, as illustrated in Figures 6-a, 6-b and 6-c, for typical periods of stable arrival rate in the non-spam, spam and aggregate workloads, respectively. To evaluate the sensitivity of the distribution to the period of measurement, we looked into the distribution of inter-arrival times observed in different periods. Figure 7 presents the cumulative distributions of inter-arrival times for two distinct periods, one during the day and the other during the evening, for each workload. Figure 7-a shows that non-spam e-mail arrivals are burstier during the day, with around 86% of all inter-arrival times within 5 seconds. During the evening, only 40% of non-spam inter-arrival times are under 5 seconds. On the other hand, the distributions are the same in both periods in the spam workload, as shown in Figure 7-b. Figure 7-c shows somewhat intermediate results for the aggregate workload.

Table 5 summarizes our findings. It shows the ranges of the mean and coefficient of variation of the inter-arrival times (measured in seconds) as well as the range values of the λ parameter (e-mail arrival rate) of the best-fitted exponential distribution, for all periods analyzed, in each workload. Note that λ remains roughly stable across all periods analyzed in the spam workload. In fact, the peak arrival rate is only 25% higher than the minimum. On the other hand, the non-spam arrival rates vary by as much as a factor of five across the periods analyzed. Aggregate traffic exibits somewhat lower variations. As discussed in Section 4. The inherently different nature of spam and non-spam senders may explain the significantly different traffic patterns.

Our results are in contrast with prior work, which found that the distribution of e-mail inter-arrival times at four e-mail servers is a combination of a Weibull and a Pareto distributions [13, 14]. However, like in our workloads (see Table 5), the reported coefficient of variation of their inter-arrival times was close to 1. Moreover, our results are in close agreement to other previous work which found a non-stationary Poisson process to model with reasonable accuracy SMTP connection arrivals [27].

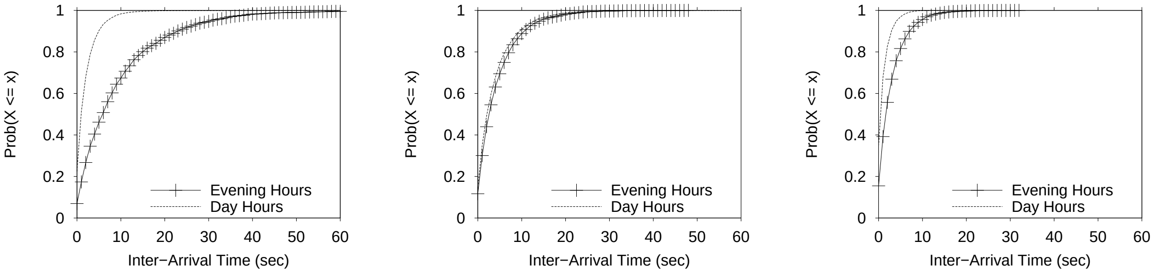
5.2 E-mail Size

We found that the distribution of e-mail sizes is most accurately approximated, both at the body and at the tail of the data, by a Lognormal distribution, in all three workloads, as illustrated in Figures 8 and 9, for a typical day. Figures 8 (a-c) show the probability density functions of the data and fitted Lognormal distributions for the non-spam, spam and aggregate workloads, respectively. Semi-log plots of the complementary cumulative distributions for the same data are shown in Figures 9(a-c). Table 6 presents the ranges of the mean, coefficient of variation and parameter values of the best-fitted Lognormal distribution for each workload, in all days analyzed. These results are consistent with those reported in previous e-mail workload characterizations [13, 14].



(a) Non-Spam (b) Spam (c) Aggregate

Figure 6: Distribution of Inter-Arrival Times



(a) Non-Spam (b) Spam (c) Aggregate

Figure 7: Sensitivity of Inter-Arrival Time Distribution to the Period Analyzed

Table 6: Summary of the Distribution of E-mail Sizes

	E-mail Sizes		Lognormal Parameters	
Workload	Mean (KB)	CV	μ	σ
Non-Spam	34 - 75	4.27 - 5.24	8.77 - 9.72	1.72 - 1.83
Spam	5 - 9	1.37 - 1.98	7.97 - 8.51	1.03 - 1.26
Aggregate	19 - 44	5.57 - 6.39	7.97 - 8.95	1.86 - 1.94

Lognormal (PDF): $p_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(ln(x)-\mu)^2}{2\sigma^2}}$.

Table 6 shows that the sizes of non-spam e-mails are much more variable and have a much heavier tail. In our workloads, approximately 83% and 61% of spam and non-spam e-mails, respectively, have sizes under 10 KB. However, whereas only 1% of all spams are sized above 60 KB, approximately 13% of non-spam e-mails have sizes above that mark. These results are consistent with those reported in [13, 14] for non-spam e-mail traffic. The impact of spam on the aggregate traffic is, thus, a decrease in the average e-mail size but an even more variable e-mail size distribution.

We draw the following insights from these results. First, in our workload, spammers typically send (a large number of) short e-mails, possibly with no attachment (content characterization is out of present scope). Second, as performed by some system administrators (including our central server administrator), e-mail size *might* be used together with other filtering techniques to improve the effectiveness of spam detection.

5.3 Number of Recipients per E-mail

This section characterizes the distribution of the number of distinct recipients per e-mail. Since this distribution is discrete, we do not apply the same fitting technique as in previous sections. Instead, like in [13, 14], we subdivide each distribution into k buckets. Each bucket is characterized with an average probability, calculated over the eight days analyzed. Jointly, these probabilities represent the distribution for number of e-mail recipients from 1 to k. For each workload, we choose a value of k so as to limit the probability of an e-mail with more than k e-mail recipients to below 0.002 [13, 14]. The values of k for the non-spam, spam and aggregate workloads are $k_{non-spam} = 8$, $k_{spam} = 16$ and $k_{aggregate} = 16$, respectively.

Figure 10 shows the cumulative distributions for the three workloads. As mentioned in Section 3.3, spams are typically addressed to a larger number of recipients. Whereas, on average, 95% of all non-spam e-mails are addressed to one recipient, only 86% of spams have a single destination. Furthermore, the distribution is heavier tailed in the spam workload, possibly due to the use of automatic tools. Since almost half of all e-mails are spams, the distribution of the number of recipients per e-mail in the aggregate workload is strongly influenced by the heavy tailed behavior observed among spams. The authors of [13, 14] found an even heavier tail in the distribution of the number of recipients per e-mail. In that study, even though 94% of all e-mails are addressed to a single recipient, 20 buckets were necessary to cover 99.8% of all e-mails.

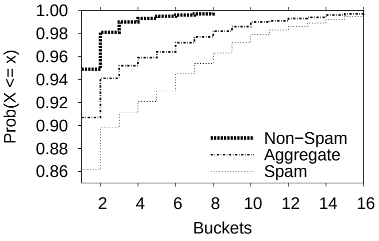
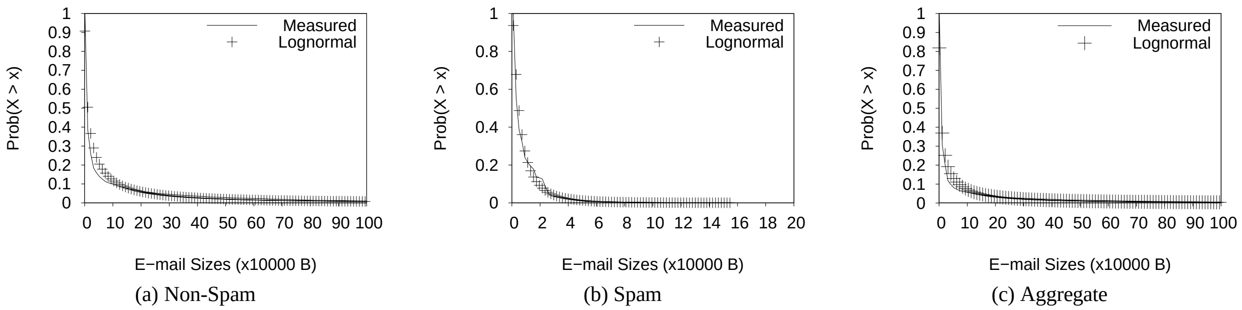
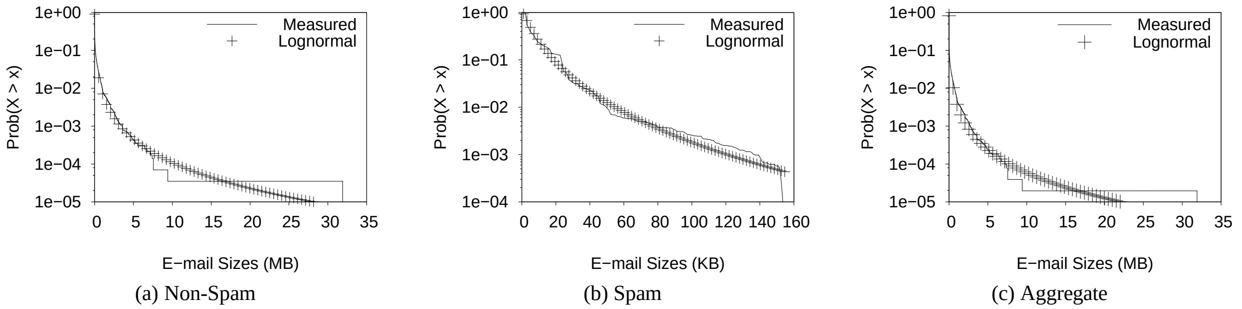


Figure 10: Distribution of Number of Recipients per E-mail



(a) Non-Spam (b) Spam (c) Aggregate

Figure 8: Distribution of E-mails Sizes (Body)



(a) Non-Spam (b) Spam (c) Aggregate

Figure 9: Distribution of E-mail Sizes (Tail)

6. ANALYZING E-MAIL SENDERS AND RECIPIENTS

This section further analyzes e-mail senders and recipients in our three workloads. Popularity of e-mail recipients and senders is analyzed in Section 6.1. Section 6.2 analyzes temporal locality among e-mail recipients and senders.

6.1 Popularity

Object popularity has been repeatedly modeled with a Zipf-like distribution (Prob(access object i) = C/i^{α}, where $\alpha > 0$ and C is a normalizing constant [33]) in many contexts, including web and streaming media [16, 17, 18, 28]. A Zipf-like distribution appears as a straight line in the log-log plot of popularity versus object rank. However, two roughly linear regions were observed in the log-log plots of some streaming media and peer-to-peer workloads [17, 18, 19, 34]. The concatenation of two Zipf-like distributions where suggested as a good model in such cases [17, 18].

In the following sections, we analyze the log-log plots of e-mail recipient and sender popularity, measured in terms of both the number of e-mails and the number of bytes received and sent. To assess the accuracy of our proposed models, we measure the R^2 factor of the linear regression [35] for each single Zipf-like distribution found. In our models, the values of R^2 are above 0.95 in all cases ($R^2 = 1$ corresponds to perfect agreement).

Section 6.1.1 analyzes recipient popularity. The distribution of sender popularity is discussed in Section 6.1.2.

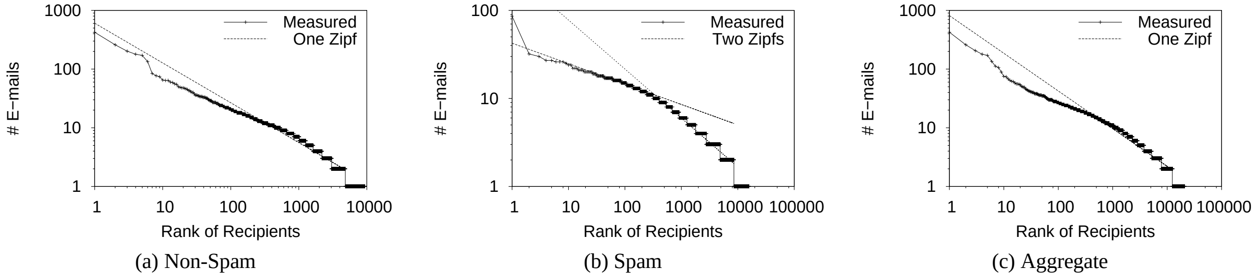
6.1.1 Recipient Popularity

This section analyzes the popularity of e-mail recipients in our three workloads. A characterization of the number of e-mails per recipient is presented first. The distribution of the number of bytes per recipient is discussed later in this section.

Number of E-mails per Recipient

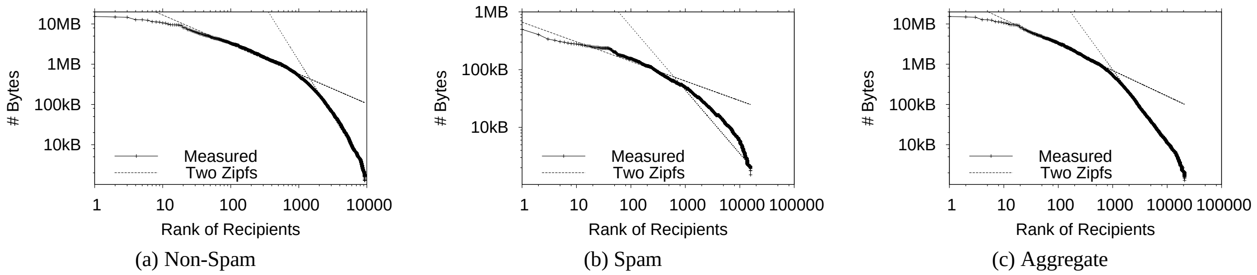
Figures 11-a, 11-b and 11-c show the log-log plots of the number of e-mails per recipient for the non-spam, spam and aggregate workloads, respectively, on a typical day. Given the large fraction of users who receive only one e-mail per day in all three workloads, we choose to characterize the number of e-mails per recipient using a combination of a fixed constant probability, for those recipients who receive only one e-mail, and a probability distribution for the remaining users.

The curves in Figure 11 present significantly different patterns for recipients of two or more e-mails. Whereas Figures 11-a and 11-c show straight lines, Figure 11-b shows two distinct linear regions in the spam workload. These results are representative of all days analyzed. Thus, for recipients of two or more e-mails per day, we model the number of e-mails per recipient with a single Zipf-like distribution, for the non-spam and aggregate workloads, and with the concatenation of two Zipf-like distributions, for the spam workload. Figures 11(a-c) show the curves for the best fitted Zipf-like distributions in each case. The roughly flat curve over the most popular spam recipients implies they receive around the same number of spams, on the particular day considered. This was true for all days analyzed, and may be explained by the larger average number of recipients per spam (section 5.3) and by the larger fraction of shared recipients among spammers (section 4.2). Again, the inherent difference between the unilateral relationship established between spammers and spam recipients and the bilateral, socially-driven relationship established between non-spam senders and recipients may incur significantly different traffic patterns.



(a) Non-Spam (b) Spam (c) Aggregate

Figure 11: Distribution of Number of E-mails per Recipient



(a) Non-Spam (b) Spam (c) Aggregate

Figure 12: Distribution of Number of Bytes per Recipient

Number of Bytes per Recipient

Figures 12-a, 12-b and 12-c show the log-log plots of the number of bytes per recipient observed on a typical day for the non-spam, spam and aggregate workloads, respectively. All three graphs show two clear linear regions, and are representative of the results found in all days analyzed. Thus, we model the number of bytes per recipient with the concatenation of two Zipf-like distributions, also shown in the graphs of Figure 12.

The discrepancy between these results and the distributions of the number of e-mails per recipient in the non-spam and aggregate workloads may be due to the high variability in non-spam e-mail sizes (see Section 5.2). Moreover, we found that the correlation between the number of e-mails and the number of bytes received by each user is typically weak, ranging between 0.18-0.27, 0.50-0.66, and 0.18-0.28 for the non-spam, spam and aggregate workloads, respectively. Thus, in each workload, the users who receive the largest number of e-mails are not necessarily the same who receive the largest volume of traffic.

6.1.2 Sender Popularity

This section characterizes sender popularity. We first present the results for the number of e-mails per sender. Analysis of the number of bytes per sender is discussed at the end of the section.

Number of E-mails per Sender

Figures 13-a, 13-b and 13-c show the log-log plots of the number of e-mails per sender, on a typical day, in the non-spam, spam and aggregate workloads, respectively. The three curves show similar behavior. As observed for e-mail recipients, there is a large number of senders that send only one message on a typical day. Moreover, the portion of the curve covering the remaining senders is well approximated with a straight line. Thus, in all workloads, we model the number of e-mails per sender with the concatenation of a constant probability, for single-message senders, and a Zipf-like distribution (shown in the Figure), for the remaining senders.

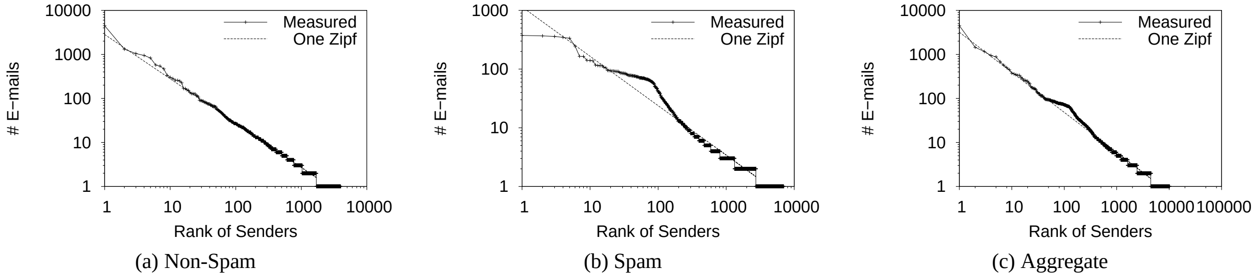
We point out that the curve in Figure 13-b somewhat flattens out over a few of the most popular spammers. However, since they represent a very small fraction of all spammers, a straight line is a reasonably good fit for the curve. Nevertheless, it is interesting to note that the fitting of the single Zipf-like distribution is more accurate for the non-spam and aggregate workloads.

Number of Bytes per Sender

A single Zipf-like distribution was found to be a good approximation of the number of bytes per sender, in all three workloads, as illustrated in Figure 14. We point out that the high variability in e-mail sizes, which might be responsible for a noticeable flat region over the recipients with the largest number of e-mails (Section 6.1.1), is less effective here because of the larger number of e-mails per sender. Furthermore, unlike observed for recipients, the correlation between the number of e-mails and the number of bytes for each sender was typically high, in the ranges of 0.68-0.88, 0.66-0.80 and 0.70-0.87, for the non-spam, spam and aggregate workloads, respectively.

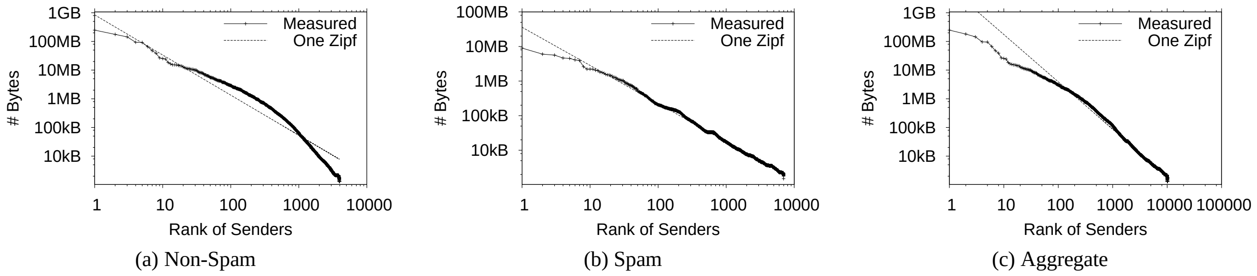
In summary, our key conclusions with respect to e-mail sender and recipient popularity are:

- The distributions of the number of non-spam e-mails per sender and recipient follow, mostly, a Zipf-like distribution. This result is consistent with previous findings that the connections between e-mail senders and recipients are established using a power law (e.g., a Zipf distribution) [28, 29].
- The distribution of the number of spams per recipient does not follow a true power law, but rather, presents a flat re-



(a) Non-Spam (b) Spam (c) Aggregate

Figure 13: Distribution of Number of E-mails per Sender



(a) Non-Spam (b) Spam (c) Aggregate

Figure 14: Distribution of Number of Bytes per Sender

gion over the most popular recipients. This may be caused by large spam recipient lists and large number of recipients shared among spammers. The number of spams per sender is reasonably well approximated with a Zipf-like distribution.

- In all three workloads, the number of bytes per recipient is most accurately modeled by two Zipf-like distributions. In the case of the non-spam and aggregate workloads, this is probably due to the high variability in e-mail size. The distribution of the number of bytes per sender is well modeled by a single Zipf-like distribution in all three workloads.

Table 7 summarizes our findings. It presents the ranges of the observed percentage of recipients/senders that received/sent only one e-mail on a typical day. It also shows the range of parameter values for the Zipf-like distributions that best fit the data for the remaining recipients/senders. For the cases where two Zipf-like distributions are the best model, Table 7 shows, for each single distribution, the total probability and percentage of recipients/senders that fall within the corresponding region of the curve as well as the value of the α parameter.

We point out that the skewed distributions of the number of e-mails and bytes per sender and per recipient suggest that sender and recipient popularity could be used to improve the effectiveness of spam detection techniques. For instance, on each day, on average, 53% of the spams and 63% of the spam bytes originate from only 3% of all strictly spam senders. Furthermore, around 40% of these spammers are among the most active throughout the eight days covered by our log. Thus, the insertion of these popular spammers into black lists could significantly reduce the number of spams accepted by the server. Similar results were observed for the senders who sends *only* non-spam e-mails, suggesting the use of white lists to avoid the overhead of scanning a significant fraction of all e-mails. Finally, the concentration of spams into a small fraction of recipients, who remain among the most popular through several days, suggests that spam detection techniques might use the e-mail destination to improve its success rate.

6.2 Temporal Locality

Temporal locality in a object reference stream implies that objects that have been recently referenced are more likely to be referenced again in the near future [30]. A previously proposed method to assess the temporal locality present in a reference stream is by means of the distribution of stack distances [30]. A stack distance measures the number of references between two consecutive references to the same objetct in the stream. Shorter stack distances imply stronger temporal locality.

This section analyzes temporal locality among recipients and among senders in our three workloads. We start by creating a set of e-mail streams, one for each workload and day analyzed. Each stream preserves the order of e-mail arrivals in the corresponding workload and day. To assess temporal locality among recipients, each e-mail in a stream is replaced with its recipient list, creating, thus, a recipient stream. The distribution of stack distances in the recipient stream is then determined. Similarly, to assess temporal locality among senders, each e-mail is replaced with its sender and the distribution of stack distances is determined.

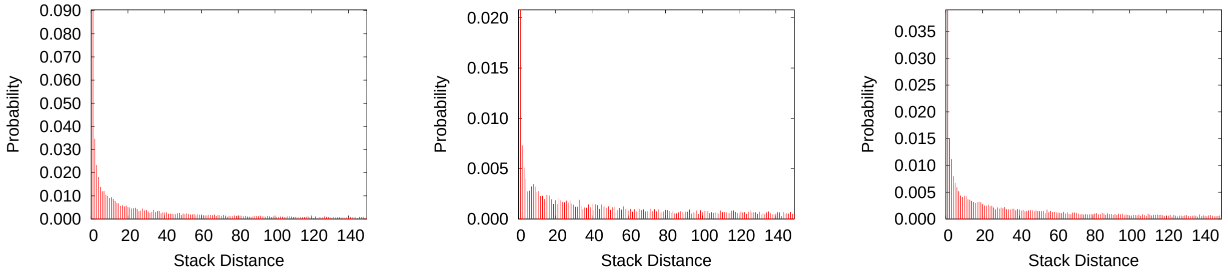
Section 6.2.1 analyzes temporal locality among recipients. Temporal locality among senders is discussed in Section 6.2.2.

6.2.1 Temporal Locality Among Recipients

Figures 15-a, 15-b and 15-c show histograms of recipient stack distances, for distances shorter than 150, observed on a typical day in the non-spam, spam and aggregate workloads, respectively.

Table 7: Summary of Distributions of Recipient and Sender Popularity

	Workload	Popularity Metric	% receive/send one e-mail/day	1^{st} Zipf			2^{nd} Zipf		
				% Data	Prob.	α	% Data	Prob.	α
Recipient	Non-Spam	# E-mails	48-62	100	1	0.521-0.678			
		# Bytes		3-21	0.05-0.26	0.622-0.838	79-97	0.74-0.95	1.674-3.204
	Spam	# E-mails	29-59	2-3	0.05-0.08	0.218-0.295	97-98	0.92-0.95	0.468-0.641
		# Bytes		4-62	0.05-0.66	0.340-0.561	38-96	0.34-0.95	1.100-3.871
	Aggregate	# E-mails	29-49	100	1	0.947-0.969			
		# Bytes		3-85	0.05-0.88	0.637-1.317	15-94	0.13-0.95	1.852-8.618
Sender	Non-Spam	# E-mails	54-68	100	1	0.993-1.251			
		# Bytes		100	1	1.72-2.08			
	Spam	# E-mails	55-67	100	1	0.781-0.996			
		# Bytes		100	1	0.915-1.192			
	Aggregate	# E-mails	53-61	100	1	0.937-0.987			
		# Bytes		100	1	1.185-1.775			



(a) Non-Spam

(b) Spam

(c) Aggregate

Figure 15: Histograms of Recipient Stack Distances

The complementary cumulative distributions of recipient stack distances, measured on that same day, are shown in Figure 16. Note the log scale on the y-axis in Figure 16.

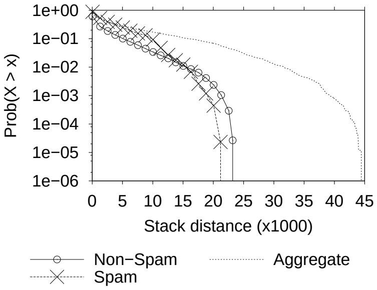


Figure 16: Complementary Cumulative Distributions of Recipient Stack Distances

We draw the following conclusions. First, there is a higher probability of very short stack distances, and thus, stronger temporal locality, in the non-spam workload. Second, the distribution of stack distances has a slightly heavier tail for non-spam recipients than for spam recipients (see discussion below). Finally, the impact of spam on the aggregate traffic is clear: temporal locality among recipients is significantly reduced, evidenced by an even heavier tail in the stack distance distributions. A summary of best-fitted distributions for recipient stack distances in each workload is presented in Table 8. Note the smaller mean and larger coefficient of variation in the non-spam workload, in agreement with our discussion above.

In search for an explanation for the significantly different temporal locality observed among spam and non-spam e-mail recipients, we defined two regions in each stack distance distribution: the head and the tail. The head (tail) consists of the shortest (largest) stack distances such that the total probability of the region does not exceed $p = 0.2$. We then defined two sets of recipients, one including all recipients for which the stack distances in the head were observed, and the other containing the recipients for which the stack distances in the tail were observed. We made three observations. First, the sets are mostly disjoint, in both spam and non-spam workloads. Second, the number of distinct recipients in the head of each distribution is a significant fraction (over 30%) of the number of daily recipients, in both workloads. Third, whereas the number of recipients in the tail of the non-spam distribution is significant, the number of recipients in the tail of the spam distribution corresponds to only 4% of daily spam recipients.

Table 8: Summary of the Distributions of Recipient Stack Distances

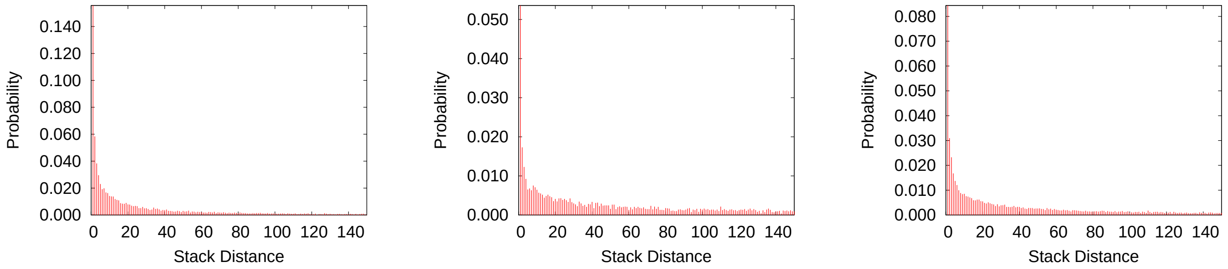
Workload	Mean (x1000)	CV	Weibull Parameters	
			α	β
Non-Spam	0.7-1.9	2.0-2.3	0.08-0.12	0.35-0.41

Workload	Mean (x1000)	CV	Gamma Parameters	
			α	β
Spam	2.2-3.5	1.1-1.6	0.36-0.54	4877-23741
Aggregate	3.1-5.4	1.4-1.9	0.29-0.37	8465-47657

Gamma (PDF): $p_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{\frac{x}{\beta}}$;

Weibull (PDF): $p_X(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^\beta} I_{(0,\infty)}(x)$

These findings, jointly, imply that there are (at least) two distinct and non-negligible sets of non-spam recipients. These sets correspond to two classes of e-mail users who exist in the real world: those who make intense use of e-mail for communication and, thus, receive bursts of e-mails from their peers, mostly during the day, when they are active, and those who send, and thus, receive e-mail only sporadically. These sets are not clearly defined among spam



(a) Non-Spam

(b) Spam

(c) Aggregate

Figure 17: Histograms of E-mail Sender Stack Distances

recipients, probably because the transmission of a spam is driven by the spammer, which acts mostly independently of the recipient's intimacy with e-mail systems.

6.2.2 *Temporal Locality Among Senders*

The histograms of e-mail sender stack distances observed on a typical day, for the non-spam, spam and aggregate workloads, are shown in Figures 17-a, 17-b and 17-c, respectively. Figure 18 the corresponding complementary cumulative distributions. As observed for e-mail recipients, there is a higher probability of very short and very large stack distances for non-spam e-mail senders. The distribution for the aggregate workload has an even heavier tail, implying a significant reduction on temporal locality among e-mail senders due to spam. A summary of the best-fitted distributions for e-mail sender stack distances is given in Table 9.

7. CONCLUSION AND FUTURE WORK

This paper provides an extensive analysis of a spam traffic, uncovering characteristics that significantly distinguish it from traditional non-spam traffic and assessing how the aggregate traffic is affected by the presence of a large number of spams.

Our characterization, based on the information available on the e-mail headers, revealed that e-mail arrival process, e-mail sizes, number of recipients per e-mail, popularity and temporal locality among recipients are some key workload aspects where spam traffic significantly deviates from traditional non-spam traffic. We believe that such discrepancies are consequence of the inherently different nature of e-mail senders in each traffic. Traditional e-mail senders are usually human beings who use e-mails to interact or socialize with their peers. Spammers typically use automatic tools to generate and send their e-mails to a multitude of "potential", mostly unknown, users.

To the best of our knowledge, this was the first effort towards a more fundamental understanding of the determinant characteristics of spam traffic. In other words, it provides a first step towards the identification of a spam *signature*, which can drive the design of more robust spam detection techniques. Research directions we intend to pursue in the future include validation of our results over time and networks, characterization of e-mail content, and further analysis of the relationship between spammers and their recipients and of spammer behavior, in general.

Acknowledgments

The authors would like to thank Fernando Frota, from CECOM/UFMG, for the helpful and inspiring discussions. They are also thankful to Marcio Bunte and Fernando Frota for providing the logs analyzed in this paper.

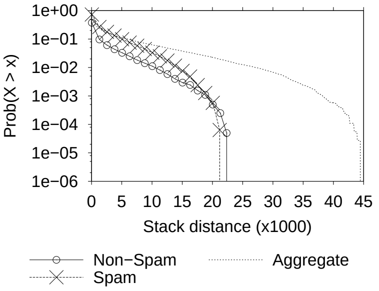


Figure 18: Complementary Cumulative Distributions of E-mail Sender Stack Distances

Table 9: Summary of the Distributions of E-mail Senders Stack Distances

Workload	Mean	CV	Weibull Parameters	
			μ	σ
Non-Spam	287-644	3.35-3.78	4.34-6.63	1.58-1.70
Spam	960-1567	1.88-2.51	5.52-7.80	1.23-1.42
Aggregate	1403-2189	2.32-3.23	6.03-7.91	1.36-1.56

8. REFERENCES

[1] N. C. Paul and C. S. Monitor, "New strategies aimed at blocking spam e-mail," http://newsobserver.com/24hour-/technology/story/655215p-4921708c.html.

[2] M. Nelson, "Anti-spam for business and isps: Market size 2003-2008. ferris research - analyzer information service, report.," April 2003.

[3] D. Fallows, "Spam: How it is hurting e-mail and degrading life on the internet," Tech. Rep. 1100, PEW Internet & American Life Project, October 2003.

[4] R. D. Twining, M. M. Willianson, M. Mowbray, and M. Rahmouni, "Email prioritization: Reducing delays on legitimate mail caused by junk mail," in *Proc. Usenix Annual Technical Conference*, Boston, MA, June 2004.

[5] MAPS Mail Abuse Prevention System Home Page, "Getting off the maps rbl," http://mail-abuse.org/rbl/getoff.html.

[6] E. Harris, "The next step in the spam control war: Greylisting," http://projects.puremagic.com/greylisting/, April 2004.

[7] L. F. Cranor and B. A. LaMacchia, "Spam!," *Comm. of the ACM*, vol. 41(8), pp. 74–83, August 1998.

[8] H. P. Baker, "Authentication approaches," in *56th IETF Meeting*, San Francisco, CA, March 2003.

[9] H. P. Brandmo, “Solving spam by establishing a platform for sender accountability,” in *56th IETF Meeting*, San Francisco, CA, March 2003.

[10] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, “A bayesian approach to filtering junk e-mail,” in *Proc. AAAI Workshop on Learning for Text Categorization*, Madison, WI, July 1998.

[11] F. Zhou, L. Zhuang, B. Zhao, L. Huang, A. Joseph, and J. Kubiatowicz, “Approximate object location and spam filtering on peer-to-peer systems,” in *Proc. Middleware*, Rio de Janeiro, Brazil, June 2003.

[12] S. Atkins, “Size and cost of the problem,” in *56th IETF Meeting*, San Francisco, CA, March 2003.

[13] L. Bertolotti and M. C. Calzarossa, “Models of mail server workloads,” *Performance Evaluation*, vol. 46, no. 2-3, pp. 65–76, September 2001.

[14] L. Bertolotti and M. C. Calzarossa, “Workload characterization of mail servers,” in *Proc. SPECTS 2000*, Vancouver, Canada, July 2000.

[15] M. Arlitt and C. Williamson, “Web server workload characterization: The search of invariants,” in *Proc. 1996 Sigmetrics Conference on Measurement of Computer Systems, ACM*, Philadelphia, PA, May 1996.

[16] E. Veloso, V. Almeida, W. Meira, A. Bestavros, and S. Jin, “A hierarchical characterization of a live streaming media workload,” in *Proc. ACM Workshop on Internet Measurement*, Marseille, France, November 2002.

[17] J. M. Almeida, J. Krueger, D. L. Eager, and M. K. Vernon, “Analysis of educational media server workloads,” in *Proc. NOSSDAV*, Port Jefferson, NY, June 2001.

[18] C. Costa, I. Cunha, A. Borges, C. Ramos, M. Rocha, J. Almeida, and B. Ribeiro-Neto, “Analyzing client interativity in streaming media,” in *Proc. 13th World Wide Web Conference*, New York, NY, May 2004.

[19] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan, “Measurement, modeling, and analysis of a peer-to-peer file-sharing workload,” in *Proc. 19th ACM Symposium on Operating Systems Principles*, Bolton Landing, NY, October 2003.

[20] C. Dewes, A. Wichmann, and A. Feldmann, “An analysis of internet chat systems,” in *Proc. ACM SIGCOMM Conference on Internet Measurement*, Miami Beach, FL, October 2003.

[21] “Exim internet mailer home page,” http://www.exim.org.

[22] “Amavis - a mail virus scanner home page,” http://www.amavis.org.

[23] “Trend micro home page,” http://www.trendmicro.com.

[24] “Spamassassin home page,” http://www.spamassassin.org.

[25] M. Delio, “Hotmail: A spammer’s paradise?,” Wired News, January 2003.

[26] B. Massey, M. Thomure, and R. B. S. Long, “Learning spam: Simple techniques for freely-available software,” in *Proc. 2003 Usenix Annual Technical Conference*, San Antonio, TX, June 2003.

[27] V. Paxson and S. Floyd, “Wide area traffic: The failure of poisson modeling,” *IEEE/ACM Trans. on Networking*, vol. 3(3), June 1995.

[28] H. Ebel, L. Mielsch, and S. Bornhold, “Scale-free topology of e-mail networks,” *Physical Review E*, vol. 66 (035103), September 2002.

[29] M. E. Newman, S. Forrest S., and J. Balthrop, “Email networks and the spread of computer viruses,” *Physical Review E*, vol. 66, September 2002.

[30] V. Almeida, A. Bestavros, M. Crovella, and A. Oliveira, “Characterizing reference locality in the www,” in *Proc. IEEE Conference on Parallel and Distributed Information Systems*, Miami Beach, FL, December 1996.

[31] E cadastro Home Page, “Lista de e-mails para venda,” http://www.e-cadastro.com/.

[32] Divulga Mail, “Servios diferenciados para mala direta por e-mail,” http://www.divulgamail.vze.com/.

[33] G. K. Zipf, *Human Behavior and the Principle of Least-Effort*, Addison-Wesley, Cambridge. MA, 1949.

[34] L. Cherkasova and G. Ciardo, “Characterizing temporal locality and its impact on web server performance,” in *Proc. Int’l Conf. on Computer Communications and Networks*, Las Vegas, NV, October 2000.

[35] K. S. Trivedi, *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*, John Wiley & Sons, New York, NY, 2001.