On the recent use of email through traffic and network analysis

The impact of OSNs, new trends, and other communication platforms

Giuseppe Aceto and Antonio Pescapè
Dipartimento di Informatica e Sistemistica
University of Napoli Federico II (Italy)
{giuseppe.aceto,pescape}@unina.it

ABSTRACT

Since the late 1971 – when Ray Tomlinson invented Internet-based email and sent the first message on ARPANET – email technology has evolved a lot, and nowadays it is one of the most widely used applications on the Internet. Despite this primacy, during the last years other ways to exchange messages have been used by Internet users (e.g. Instant Messaging, Social Networks, microblogs, etc.). In this paper we propose a methodology based on heterogeneous data sources to analyze the amount of traffic associated with emails in order to gain knowledge on the use of email by Internet users in the last years. We consider real traffic traces that are well known to the research community as well as locally captured, and discuss them in the light of other related phenomena: social networks adoption, online advertising trends, abusive email spreads, etc..We discuss the trend of email traffic in the last 10 years and we provide explanations related to the impact, on the email usage, of the utilization of other communication platforms. This work represents a first step towards a framework in which to analyze the trend of the email traffic and the associated phenomena as well as the understanding of the upcoming novel communications behavior of Internet users.

1. INTRODUCTION

In the last two decades email has evolved in both the interface (User Agents) and the message handling subsystem (Message Transfer Agents) [1] and it has been successful in displacing other types of traditional communication such as hard mails, voice calls, and FAX. As a consequence, the amount of traffic associated with email shows an increasing trend since the 90's, when – together with web browsing – email was one of the primary means of data exchange over the Internet.

However, despite the continuous growth of abusive emails ("spam") and the increasing number of "digital users", in the last few years the trend does not show the expected increasing characteristic. It seems that while for the Web we have assisted to a shift in the usage paradigm (from static content received from "passive users" towards dynamic content provided by "producers", e.g. YouTube or Flickr), another interesting and still not well investigated shift is on the run for the exchange of messages among people: *Social Networks* (e.g. Facebook, MySpace, Linkedin, Xing, Plaxo), *Instant Messaging* (e.g. AIM, MSN, Skype, Google Talk, ICQ), *microblogging* (e.g. Twitter, Jaiku, Pownce), and *novel integrated communication platforms* (e.g. Google portal, encompassing email, instant messaging, document sharing, audio and video communications) nowadays complement – and sometimes replace – email, especially for short messages.

1.1 Our Contribution

The literature on email and on the related traffic can be split into four main categories:

1. traffic characterization and modeling [2, 3, 4, 8];
2. models for study the evolution of email networks [9, 10, 11];
3. spam analysis and detection [12, 13, 14];
4. study of consumer emails for marketing analysis [15, 23].

To the best of our knowledge, there is no substantial work on (i) understanding the trend in the use of email in the last 10 years; (ii) analyzing the use of email with respect to the growth of Internet users, the percentage of spam, the increasing use of Online Social Networks (OSNs), etc.

In this paper, we propose a simple methodology based on heterogeneous data sources and we highlight and discuss the changes in email usage and provide some motivations using real data. By adopting a ***multi-variables*** and ***multi-sources*** approach, we perform a ***multiple-viewpoints*** analysis in order to find possible interactions among strictly related phenomena. Using a very large number, **20**, of different data sets and the outlined approach, we carefully analyze the behavior of email traffic linked to other phenomenon over a large time period. To the best of our knowledge, the novel contributions of our study are the following:

- we found, over a large set of traffic traces, evidences of a small decreasing trend in the use of email adopting SMTP (Simple Mail Transfer Protocol) and related to TCP port 25 and we provide for the first time in literature experimental results supporting the claim;
- we found that the average size of the exchanged emails (adopting SMTP and related to TCP port 25) is increased over the time (this is a symptom of the fact that short messages are often sent using other communication platforms) and we provide the analytical statistical fitting of the PDF of the packet sizes;

- adopting the proposed *multiple-viewpoints, multi-variables and multi-sources* approach, we are able to carefully analyze the trend we depicted with respect to other inherent phenomena: Internet users base variation, spam, social networks adoption, Internet advertising revenues and budget plans;
- Internet users are even more using OSNs and systems like Twitter to exchange short messages.

We present the methodology, the data it is applied to, and considerations on both in Sec. 2; then in Sec. 3 we describe the results of the application of the presented approach to the available data, and we discuss the findings; finally in Sec. 4 we draw our final remarks and future improvements for the proposed approach.

2. EMAIL TRAFFIC ANALYSIS

In this section, we present the methodology we adopted (Sec. 2.1); then we briefly describe the data sets used in this work (Sec. 2.2); finally we (Sec. 2.3) discuss the issues we faced with the data.

2.1 Methodology

The subject of the present analysis is the inference of long term usage analysis of email communications using SMTP and related to TCP port 25, with a total time span of 10 years, based on network traffic traces and integrating information from related phenomena; these complementary points of view contribute to better understanding the evolution of email seen as part of a single, but complex, online communication ecosystem taking advantage of the well known [4, 5, 6, 7] properties of time and space invariance shown by Internet communications.

Traditional approaches rely on the analysis of network data traces. Ideally, having access to full payload data traces would have allowed for an analysis at application level, but this poses two main problems: the first is privacy-related, as application level capturing would expose varying degree of sensitive information about users; as a consequence, for ethical and legal reasons, full payload traces are generally not publicly available; a second problem is that the mere amount of data to be kept for long timespan captures makes it unfeasible from a practical and economic point of view. This has lead to the use of header-only and anonymized traffic traces in network traffic studies. Our work starts with an analysis performed on well known traffic traces reporting only network and transport headers. The analysis at network and transport level by itself is not able to provide enough information for detecting changes in email usage patterns, but can be profitable if enough context information is provided. We looked for such context in publicly available data on phenomenons related to email usage and integrated that information in our analysis. We also provided motivations for the detected phenomenon and supported them with economic-related sources. At the best of our knowledge, it is the first time in literature that an integrated ***multiple-viewpoints***, ***multi-variables*** and ***multi-sources*** approach and the derived results are presented in the field of email traffic monitoring and analysis.

Many different factors could affect the amount of email traffic that is exchanged on a link: the most obvious one, that is what we want to detect, is the number and size of email messages exchanged by users. However, as we aim at inferring human users' behavior, a major issue is related to machine-generated content, and specifically undesired messages including SPAM, email worms, and malicious attachments not related with willful human activity: to track the overall incidence of this aspect, we included in the analysis the data from security field. Another factor that affects the amount of email traffic is clearly the growth of the number of users. To take into account this variable, we compared the trend of capture traffic against the statistics on the Internet user base.

These heterogeneous sources have been juxtaposed on the same time reference and subject to integrate analysis. This let us derive a number of findings, for which we provided possible justifications. In order to support such claims, we included in the discussion external data regarding OSNs usage and advertising trends. The characteristics of used data and issues related with their adoption and interpretation are the subject of the following sections.

2.2 Data Sets

The presented methodology has been applied leveraging **20** different datasets, coming from different kinds of sources and with heterogeneous nature; they can be grouped in five categories: network traffic traces, OSNs usage logs, Internet usage statistics, malicious or abusive email reports, and surveys about online advertising. All data considered in this paper are reported in Tab. 1, highlighting the time span and the granularity of the data collection; in the column "granularity" the label *continuous* marks data with associated timestamping, while *snapshot* means that collected data have no explicit timestamping, so they represent a snapshot at the time of the collection.

Traffic Traces

We used traffic traces provided by the MAWI-WIDE project [16]. Such traces are widely adopted by the research community: analyses of different aspects of the network traffic in MAWI traces are present in literature, e.g. focusing on anomalies [18, 19] and long-range dependency [20]. The MAWI-WIDE project makes available traffic traces of 15 minutes for each day of the year since 2000 from several links: in this way we have data related to a long period of email traffic captured from the same link and at the same time. We considered the SMTP traffic related to TCP port 25. We used the traffic traces captured over a trans-Pacific line (samplepoint-B and samplepoint-F at [16]). MAWI traffic traces are often used by researchers in the networking community also because of their nature: they are related to transoceanic links and then they provide a high degree of generalization for the analyses carried out using the data extracted from them.

We also used traffic traces collected at a link at 200Mbps connecting the University of Napoli "Federico II" network to the rest of the Internet. This traffic is related to TCP port 25 generated by clients inside the network of University of Napoli, Federico II - UNINA - reaching the outside world (i.e. src host from UNINA and dst port tcp 25 OR dst host from UNINA and src port tcp 25); the capture lasts one hour (from 11:00 to 12:00 of September 5th, 2005) [48, 49]. Another trace from the same link has been captured from 12:00 to 13:00 of July 13th 2010.

Time Span	Granularity	Kind	Source
2001/01/01 - 2006/06/30	daily, 14:00-14:15	pcap, 96bytes IPv4	mawi samplepoint B [16]
2006/08/24 - 2006/09/03	daily, 14:00-14:15	pcap, 96bytes IPv4	mawi samplepoint F [16]
2006/10/03 - current	daily, 14:00-14:15	pcap, 96bytes IPv4	mawi samplepoint F [16]
2003/12/1 - 2003/12/15	continuous	ERF, anonymized, zeroed payload	WITS:AucklandVIII [17]
2005/09/05	single day, 11:00-12:00	pcap, anonymized, 96bytes IPv4	UNINA [48]
2010/07/13	single day, 12:00-13:00	pcap, anonymized, 96bytes IPv4	UNINA [48]
2004/02 - 2006/03	continuous	messages and pokes headers	Facebook [28]
2006/03/31	snapshot	friends list per user	Facebook [28]
2006/09/26 - 2009/01/22	continuous	wall posts	New Orleans Network (Facebook) [27]
2008/12/29 - 2009/01/03	snapshot	public profile data	New Orleans Network (Facebook) [27]
2009/03/26 - 2009/04/06	continuous	http requests through aggregator	Orkut, MySpace, Hi5, LinkedIn [30]
2009/04/10 - 2009/04/17	snapshot	public profile data.	Orkut [30]
2001 - 2009	yearly	survey on Internet advertising revenues	U.S. adverstising companies [43, 44]
2009	yearly	survey on email marketing budget plans	U.S. media companies [41]
2009 - 2010	yearly	surveys on marketing budget plans	U.S. media companies [40, 42]
1995 - 2010	quarterly - yearly	Internet usage statistics	various, worldwide [21]
2007 - 2008	yearly	access statistics to online services	worldwide [45]
June 2009; June 2010	continuous	access statistics to online services	USA [46]
last quarter 2005 - 2008	quarterly	percentage of abusive email	various, worldwide [22]
2005 - 2009	quarterly	ratio of malicious emails	worldwide [31]

Table 1: Data sources analyzed in the present work.

OSNs Dataset
For the analysis related to OSNs, the considered datasets in [28] consist of timestamped anonymized headers of messages and pokes (content-less messages) between February 2004 and the end of March 2006, among 4.2 million of North American college and university students, provided by Facebook. The New Orleans regional Network in Facebook was crawled in [27], from December 29th, 2008 to January 3rd, 2009, gathering the topology of "friendship" network, and between January 20th and 22nd 2009 the wall history (record of timestamped public *friend-to-friend* messages) of collected users was retrieved, obtaining data with a timespan ranging from September 26th, 2006 to January 22nd, 2009. Data in [30] cover the period March 26 - April 6 2009, and refer to session-level summaries of HTTP requests, gathered though multi-sign-on OSNs aggregator providing access to Orkut, MySpace, Hi5, LinkedIn for 37,024 users; of Orkut users profile data were gathered from April 10 to 17 2009.

Internet Usage Statistics
In order to take into account the variations of usage of Internet across the analyzed time span, we considered the world Internet usage statistics from [21], that gathers data from different sources and spans from 1995 to 2010 with sampling intervals ranging from quarterly to yearly. By monitoring accesses to online services for a sample of instrumented web sites, [45] provides aggregated data related to statistical samples of population in AU, BR, CH, DE, ES, FR, IT, UK and USA) across 2007 and 2008; the same method is used [46] to obtain the breakdown of per-activity (OSNs, email, instant messaging, etc.) usage percentages, specific to U.S.A., in June 2009 and June 2010.

Malware Evolution Statistics
Data about malware and spam in emails are collected from [22] by *MAAWG (Messaging Anti-Abuse Working Group)*, an association focused on addressing various forms of messaging abuse; provided reports are compiled quarterly from aggregated data collected by member ISPs, email providers and network operators (mostly from U.S.A., but also France and Germany), on the interval from last quarter of 2005 to 2010. A closely related point of view is taken from reports by *SophosLabs* [31], the global network of researchers and analysts of a company in the field of information security products and services; referenced statistics span from 2005 to 2009, aggregated yearly.

Internet Advertising Statistics
As a source for marketing data regarding Internet advertising, we used publicly available reports by *Marketing Sherpa*, a research firm specializing in analysis of marketing trends. Namely, we referred to [40], reporting the results of a fielded survey held on August 2010, with a sample base of 935 interviews about changes in marketing budget plans, divided in 11 categories comprising "email marketing" and "social media"; we compared those results with previous year surveys [41, 42] by the same source. About Internet advertising, another source we used are publicly available reports sponsored by the *Interactive Advertising Bureau (IAB)*, an association of technology and media companies covering a significant share of online advertising market in the United States. Data were taken from [43, 44], a series of surveys conducted by *PricewaterhouseCoopers (PWC)* interviewing companies that sell advertising online, and focusing on revenues from different kinds of media (web sites, email, etc.); the reports are held quarterly since 2001, we referenced full-year aggregated data.

2.3 Issues with the data

While mixing different data sources, analyzing heterogeneous data, and inferring indirect relations can give interesting and novel insights on long term phenomena, it is important to underline several concerns and problems that could arise and that a researcher must consider when adopting an integrated ***multiple-viewpoints***, ***multi-variables*** and ***multi-sources*** approach.

2.3.1 SMTP data assumptions

In order to analyze the trend of email usage, we make the simplifying assumption that the number of exchanged IP packets carrying SMTP messages is proportional to the number of issued email messages, by a constant factor. Actually, an (E)SMTP session can imply a varying number of messages between client and server to complete a single email delivery [47], specifically the enabling of STARTTLS

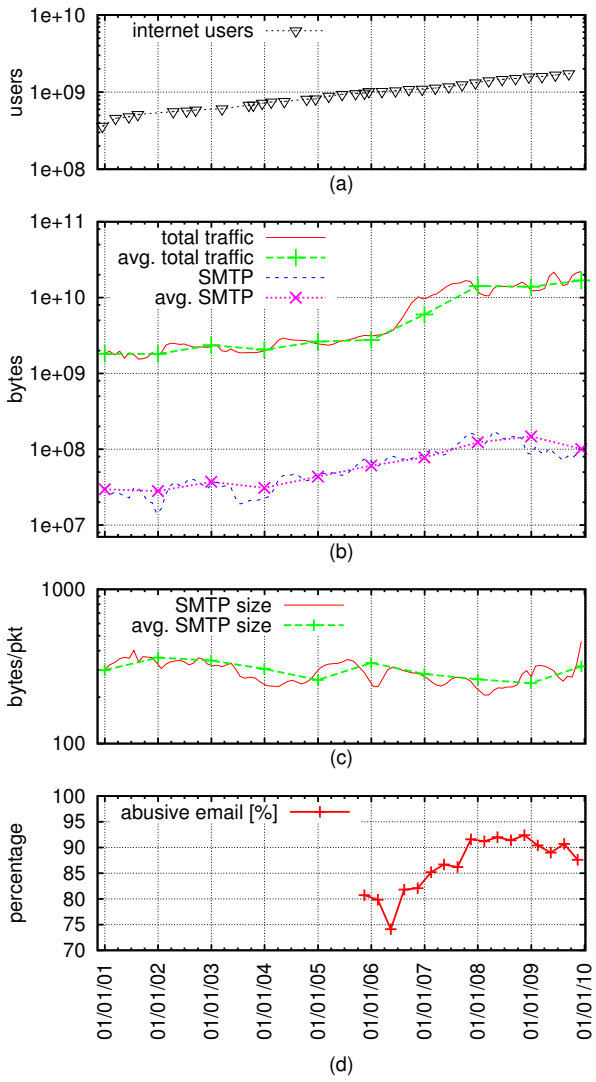


Figure 1: Synoptic time series: a) Internet users population; b) captured traffic (total traffic, SMTP traffic on TCP port 25); c) average SMTP packet size on TCP port 25; d) percentage of abusive emails. Data in a) are from [21]; b) and c) from [16]; data in d) are from [22]. All graphs share the same x axis (date). For traffic traces the thin lines are Bezier smoothing of raw data, while thick lines with points are averages on a year.

for authentication and privacy, PIPELINING, and CHUNKING extensions and the different chunk size can change the number of messages; in any case, the SMTP protocol forces a single message from sender to *not* contain more than one email, in any case. This means that the adoption and the spreading of such extensions in time lead at most to the increasing of the number of exchanged packets, the other parameters intended as constants. Another variable is the size of transport PDUs: an increasing in TCP average MSS would lead to the exchange of less packets, with increasing average size.

2.3.2 *Restricted data*

Some of the sources only provide analysis results, without disclosing the data on which those analysis have been performed, nor the detail of the methods, thus preventing verification. This raises an obvious concern when the publisher of the analysis is a stakeholder in the field that is investigated. As the phenomena we are considering do not directly affect the interests of the cited stakeholdes, and being the selected sources well known and authoritative in their respective fields, we deemed the reported data as reliable for the sake of our analyses.

2.3.3 *Different geographic zones*

Heterogeneity in the geographic zones is an obstacle to the analysis of the phenomenon, as different cultures and different economies could present local causes to the presence (or absence) of corroborating elements. Moreover, as the only criterion to correlate such data is chronological, the presence of some sort of lag in adoption of technologies, habits, specific applications, could introduce a time shift that further impairs the analysis. All considered data includes at least U.S.A., adding consistency to the analysis, but for the rest has weak global coverage, making this approach and results not straightforward generalizable to phenomena that do not show spatial invariance.

2.3.4 *Different time spans*

Heterogeneity in the time intervals and in the granularity is an obstacle to an easy temporal correlation analysis, leaving blind spots. This has necessarily focused our analysis on the time intervals for which the more sources were available.

2.3.5 *Opinion Mining*

Some of the data ([40, 42, 41]) are the results of opinion mining, and refer to budget *plans* for the year subsequent to that of the survey: differently from actual revenues ([43]) these involve also *expectations* and forecasting, that could be influenced by factors external to the ones of our interest. Nevertheless, specific budget allocation is the objective mirror of the perceived importance of email marketing for the enterprise, and finally is very likely to have impact on the email volume, therefore we decided to take into account this factor in our analysis.

2.3.6 *Loosely correlated multiple causes*

The depletion phenomenon we describe can be the result of many non-synchronized, loosely-correlated causes, possibly with non-monotonic contribution: their interaction could mask each other; this is not a problem as long as we are interested on overall "ensemble" result, not in identifying the relative contribution of each cause.

The issues reported and discussed above (Sec. 2.3.1-2.3.6) are to be considered as a first proposal of checklist-like warnings and boundaries, and part of a guideline on ***multiple-viewpoints***, ***multi-variables*** and ***multi-sources*** analysis of a real-world, lively complex system, for which relying on different and heterogeneous data sources and sets for inferring indirect relations is the only way we have to derive interesting and novel insights on long term phenomena. More precisely, we believe that using the data and the described methodology permits to analyze macroscopical and long term phenomena like email usage analyzed in this paper. In addition, the temporal and spatial invariance of both the Internet traffic properties and Internet phenomena - many times found and described in literature [4, 5, 6, 7] - guarantee on the applicability and the validity of our assumptions as well as on the approach, methodology and results.

3. EXPERIMENTAL RESULTS

In Fig. 1b the total traffic (upper line, green '+') and the SMTP traffic on TCP port 25 (lower line, purple 'x') show different variable trend. We also report the Internet users (IU) trend (Fig. 1a, black '∇') to have a reference for the relation between the growth of users and the number of exchanged messages using SMTP on TCP port 25. The Internet users trend shows the steady and well known power-law evolution (in the logarithmic scale it shows a good approximation of a line). The total traffic shows two phases: approximatively parallel (in logarithmic scale) to the IU trend, from 2001 to 2006 and from 2008 to the end of 2009; dramatically increasing from 2006 to 2008 (growing much faster than IU), which could be attributed to the wide use of social networks and user-generated content. SMTP traffic grows at a slower pace than IU from 28MB/15' in 2001 up to 43MB/15' in 2004. Then a steady power-law ramps-up towards 148MB/15' at about half 2008 (faster than IU, but not as the explosion of total traffic in 2006-2008), then it shows an evident decrease down to 101MB/15' in the end of 2009 (see Tab. 2). We argue that the change in the trends, not reflecting IU, is due to a diffuse change of usage patterns in the Internet:

- preexisting users do different tasks (or in different ways) than before; for example, there is an increasing use of webmail with respect of using email clients (thus using HTTP/HTTPS instead of SMTP for carrying email messages and therefore port 80 and 443) [23];
- preexisting users can configure a client to use a secure access (e.g. on port 465);
- newcomers introduce and spread new usage patterns, mainly using social networks and Instant Messaging platforms (*consumerization of Internet applications*) [24, 25].

In particular, in the last two years the SMTP traffic shows a countertrend with respect to both total traffic and IU: in our opinion this phenomenon advocates for the transition from email using SMTP to other forms of Internet-mediated personal interactions (see Sec. 3.1). Finally, it is worth noting that our claims are corroborated by the increasing trend (averages on year from 79% in 2006 up to 91% in 2008) of the spam (see Fig. 1d, red '+') that contributes to further decreasing the actual legitimate email usage.

Year	traffic (MB)	StdDev (MB)	Samples	Min (MB)	Max (MB)
2001	28.08	13.41	352	0.51	100.16
2002	37.18	20.64	358	1.20	166.67
2003	30.95	21.24	350	4.83	233.76
2004	43.68	22.01	364	9.71	136.81
2005	60.98	28.53	340	11.91	216.37
2006	78.11	30.74	276	8.26	185.03
2007	123.50	57.44	350	40.59	404.54
2008	147.53	66.69	358	12.14	464.38
2009	100.57	59.07	343	9.58	424.39

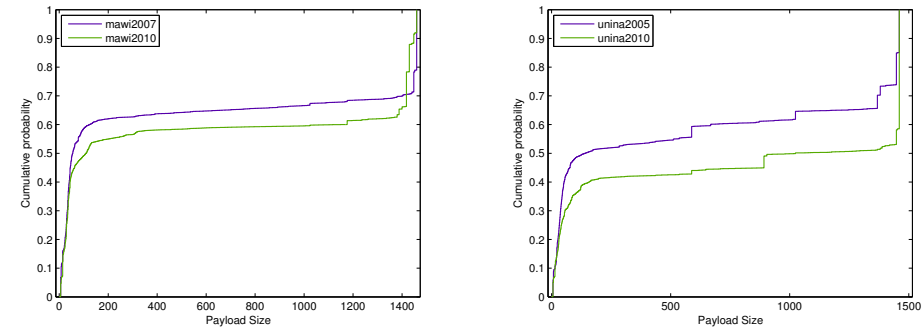
Table 2: Statistics of SMTP traffic (from [16]).

In the time interval in which SMTP traffic decreases we can observe an increasing trend of the average SMTP packet size (from 256 to 316 bytes per packet, see Fig. 1c, green '+' line): it testifies an even stronger decreasing usage of email using SMTP on TCP port 25, as SMTP traffic is built up on fewer but larger packets (see Sec. 3.1). To provide more details, in Fig. 2(a) we show the Empirical Cumulative Distribution Functions of two MAWI traces (one from 2010, on Tuesday July 13th, and the other from the closest date in 2007[1] with the same weekday: July 10th). The graph shows a significant prevalence of small-sized packets for the older trace, in which there is an increase of the average payload size from 527.25 to 613.02 bytes, with variance equal to $2.81 \cdot 10^5$ and $4.45 \cdot 10^5$ respectively. This trend is confirmed using a completely different couple of traffic traces. In Fig. 2(b) we show the Empirical Cumulative Distribution Functions of the two traces (one from 2010 - the same day of MAWI trace - and the other collected in 2005 on September 5th) captured at University of Napoli [48]: again, small size packets are less frequent in the new trace (increasing from 613.26 to 809.32 bytes, with variance of $4.15 \cdot 10^5$ and $4.51 \cdot 10^5$ respectively), again supporting a transition towards bigger messages.

In order to shed lights on this trend we analyze more in depth the considered traces, whose Empirical Probability Distribution Functions are shown in form of histogram in Fig. 2(c) and in Fig. 2(d) for MAWI and UNINA traces, respectively. The size of the bin is calculated according to the *Scott's rule* [51]. All traces exhibit two peaks: we divided each distribution in two intervals, following the *Maximum Fisher Distance* criterion [50]. The results of the partitioning criterion for the four traces are shown in Tab. 3. It can be noticed that even if the cutting point shifts towards the bigger values only for the MAWI traces, for both cases the fraction of payloads with size falling in the interval $[cut, 1460]$ (see column p_2) increases, again confirming our claim. By using the partitions defined according to the *Maximum Fisher Distance*, each trace has been modeled by means of parametric PDFs, separately fitting each subpopulation ad evaluating the best fit according to the λ^2 *metric* [52]. As reported in Table 4, we found that for all traces[2] the upper subpopulation can be modeled with a Normal, with mean $1382.13, 1396.83, 1407.55, 1407.78$ and standard deviation $148.45, 169.12, 140.99, 87.96$ for UNINA 2005, UNINA

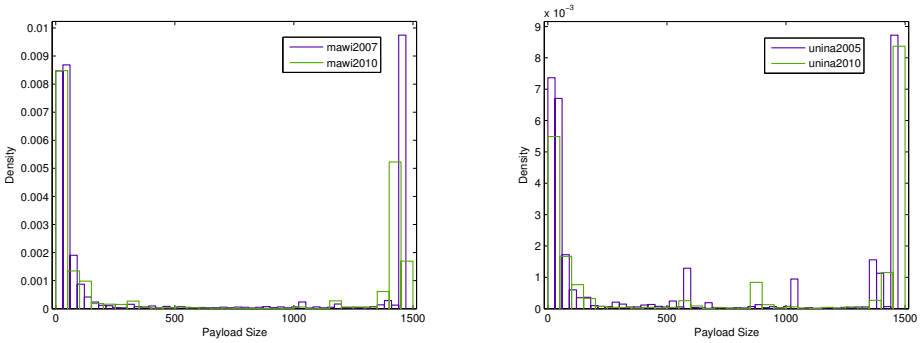
[1]We have found similar results for the same day of the year, July 13th 2007.

[2]For MAWI 2007 upper subpopulation, the Normal is second best, with a λ^2 increased by 0.08 over the one of the Weibull with scale parameter 1447.99 and shape parameter 24.91.



(a) Empirical CDF of MAWI PS

(b) Empirical CDF of UNINA PS



(c) Empirical PDF of MAWI PS

(d) Empirical PDF of UNINA PS

Figure 2: Empirical CDF and Empirical PDF TCP Payload Size (PS) for SMTP (port 25): traffic traces UNINA (2005/09/05 and 2010/07/13) and MAWI (2007/07/10 and 2010/07/13).

Trace	Cut	FD	μ_1	${\sigma_1}^2$	p_1	μ_2	${\sigma_1}^2$	p_2
UNINA 2005/09/05	751.3340	59.2630	109.1434	3.0824e+004	0.6040	1382.1	2.2037e+004	0.3960
UNINA 2010/07/13	724.9177	72.7945	81.2909	1.7791e+004	0.4466	1396.8	2.8602e+004	0.5534
MAWI 2007/07/10	717.6333	143.4020	58.2098	8.8699e+003	0.6524	1407.6	1.9879e+004	0.3476
MAWI 2010/07/13	733.8084	219.7946	63.6402	8.5544e+003	0.5913	1407.8	7.7362e+003	0.4087

Table 3: Bimodality analysis: Cutting point, *Fisher Distance* (FD) and statistics (mean, variance, fraction of population size relative to the intervals $[0, Cut[$ and $[Cut, 1460]$, orderly) for UNINA and MAWI traffic traces.

2010, MAWI 2007, and MAWI 2010 respectively. Similar consistence is found across the lower subpopulations, that can be all[3] modeled as Lognormal, with parameters μ equal to $3.77, 3.63, 3.47, 3.57$ and σ equal to $1.31, 1.19, 1.02, 1.03$ for the traces UNINA 2005, UNINA 2010, MAWI 2007, MAWI 2010 respectively.

Finally, it is worth noticing that the change in time of adoption of SMTP extensions, as said in Sec. 2.3.1, can not held responsible for this, as it would imply a relative increase of control messages over content-related messages, leading to *smaller* average message size.

3.1 Discussion

In this paper we analyze the trend of SMTP-based email traffic in the last 10 years, compared to the increasing number of Internet users and spam emails. According to the nature of scale-free behavioral network shown by email [26], we would expect a growing usage trend at least analogous to the one of its user base: as shown in this paper this is not the case of SMTP traffic. In this section we discuss two main causes at the base of this phenomenon. Another motivation for the decreasing email usage to investigate is that people are getting annoyed by spam, malware (viruses, trojans, identity theft, phishing, scams, spyware, spam, adware) and distasteful contents received from emails. Hence, they would prefer other communication platforms currently affected by a lower amount of malware. This claim needs to be supported by real data and is out of the scope of this paper.

3.1.1 *Using OSNs to exchange messages*

In our opinion one of the main causes of the trend depicted in this paper is related to the use of Social Networks, Instant Messaging, and microblogging platforms to exchange private

[3]The best fit for UNINA 2005 lower subpopulation, with a λ^2 smaller by 1.5, is a Weibull with scale parameter 85.84 and shape parameter 0.73.

UNINA 2005						
	lower subpopulation			upper subpopulation		
Distribution	par 1	par 2	λ^2	par 1	par 2	λ^2
normal	109.1434	175.5677	9.765712e+000	1382.1375	148.4493	**1.591119e+001**
exponential	109.1434		1.254145e+001	1382.1375		3.544900e+001
weibull	85.8356	0.7329	**5.149531e+000**	1431.5954	18.9153	2.075107e+001
gamma	0.6625	164.7411	6.005542e+000	71.1632	19.4221	1.755544e+001
extreme	211.8663	237.8659	6.856953e+000	1433.9171	66.0986	3.077077e+001
lognormal	3.7734	1.3071	6.657950e+000	7.2243	0.1248	2.077694e+001
rayleigh	146.1785		1.550991e+002	982.9398		3.221151e+001
uniform	1.0000	751.0000	6.872852e+000	752.0000	1460.0000	2.929172e+001

UNINA 2010						
	lower subpopulation			upper subpopulation		
Distribution	par 1	par 2	λ^2	par 1	par 2	λ^2
normal	81.2909	133.3829	1.554139e+001	1396.8252	169.1197	**2.391448e+001**
exponential	81.2909		5.994306e+000	1396.8252		5.338921e+001
weibull	69.6389	0.8022	1.959193e+000	1447.7230	19.2464	2.548467e+002
gamma	0.7762	104.7335	3.522146e+000	53.1406	26.2854	2.708313e+001
extreme	163.5378	212.1713	4.945752e+000	1449.6669	61.0052	5.044954e+002
lognormal	3.6303	1.1895	**1.365945e+000**	7.2325	0.1463	2.608689e+001
rayleigh	110.4518		7.096520e+000	994.9176		4.475599e+001
uniform	1.0000	724.0000	5.714118e+000	725.0000	1460.0000	3.986205e+001

MAWI 2007						
	lower subpopulation			upper subpopulation		
Distribution	par 1	par 2	λ^2	par 1	par 2	λ^2
normal	58.2098	94.1799	1.598910e+000	1407.5539	140.9912	1.148396e+001
exponential	58.2098		1.393066e+000	1407.5539		4.575729e+001
weibull	54.5492	0.9009	9.777219e-001	1447.9965	24.9149	**1.140088e+001**
gamma	0.9806	59.3599	1.301394e+000	77.3599	18.1949	1.295428e+001
extreme	119.2087	177.7702	3.338997e+000	1449.3580	48.5768	1.271136e+001
lognormal	3.4741	1.0223	**4.431508e-001**	7.2431	0.1215	1.464041e+001
rayleigh	78.2887		4.419123e+000	1000.2716		3.946567e+001
uniform	1.0000	717.0000	7.786445e+000	718.0000	1460.0000	3.253709e+001

MAWI 2010						
	lower subpopulation			upper subpopulation		
Distribution	par 1	par 2	λ^2	par 1	par 2	λ^2
normal	63.6402	92.4896	1.522155e+000	1407.7784	87.9550	**3.748629e+000**
exponential	63.6402		8.339846e-001	1407.7784		2.260523e+001
weibull	60.3468	0.9138	6.420627e-001	1432.7721	38.7154	9.351232e+000
gamma	0.9896	64.3108	7.806842e-001	207.8376	6.7735	4.687160e+000
extreme	121.4613	165.6676	2.988674e+000	1433.5702	34.2876	1.036665e+001
lognormal	3.5692	1.0283	**4.071396e-001**	7.2474	0.0733	6.217864e+000
rayleigh	79.3864		4.343236e+000	997.3906		1.914531e+001
uniform	1.0000	733.0000	7.948789e+000	734.0000	1460.0000	1.726388e+001

Table 4: Fitting distributions and parameters for empirical PDF of PS of traces: UNINA (2005/09/05 and 2010/07/13) and MAWI (2007/07/10 and 2010/07/13).

and short messages (the diffusion of mobile devices and the consequent use of SMSs and – in last years – Mobile Instant Messaging also contributes to the phenomenon). Very often, people use these platforms instead of email to exchange short messages like "*hi, how are you?*", "*how about pizza this evening?*", for gossiping or for near-real-time news report. People is not suddenly dismissing email: the decreasing trend started because, for short messages, the quick communication loop *read-a-post / write-a-comment* is already available "on site" and "in topic", with no need of opening a specific webmail site or email client and explicitly adding recipient, subject, references and context.

In [28], analyzing Facebook data of nearly 500 North American colleges and universities, it was found that private messages (email-like, but with a single recipient) and *pokes* (a content-less message) together had an average of 0.97 messages per user per week, on a user base of $4.2 \cdot 10^6$ users, and with a heavy-tailed distribution; in the early days of its adoption, Facebook delivered an average of more than $2 \cdot 10^7$ messages or pokes per day just for the networks considered in [28]. The exchange of private, semi-private or public messages is a significant part of user activity on OSNs: as reported in [30], in Orkut, Myspace, LinkedIn and Hi5 messaging or messaging-like is in the top-five activities performed (in terms of share of HTTP requests), on traces collected in 2009. Similar results were found in [29], with messaging on Facebook summing up to more than 20% of active user requests (as opposed to automatic AJAX requests), on different traces collected in 2008. By monitoring public profiles in a regional network, in [27] is shown from September 2006 to January 2009 an increase from less than 500 to almost 2500 wall posts per month.

The growth in popularity of OSNs at the expense of email is confirmed by independent methods, e.g. according to [45], the analysis of data gathered through monitoring web portals shows that the percentage of monitored users accessing webmail in December 2008 had an increase of 2.7% with respect to a year before, while the usage of OSNs and blogs in the same period has increased by 5.4% becoming the fourth most popular browsing activity, swapping place with webmail. With the same method, in [46] it is found that the time spent monthly online in 2010 accounts for $906 \cdot 10^6$ hours on

Social Networks/Blogs versus $329 \cdot 10^6$ hours on webmail.

A confirmation of this phenomenon can be seen in the growth of the average SMTP packet length detected in 2008-2009 (cfr. Fig. 1c): a sign that longer messages or the addition of attachments are prevailing on short messages. Abusive attachments do not contribute much to this trend, as the percentage of email with malicious attachments is about two orders of magnitude lower than these variations (see Table 5, from [31]).

Year	Emails (%)
2005	2.27
2006	0.296
2007	0.110
2008	0.140

Table 5: Percentage of emails with malicious attachments (average), from [31].

3.1.2 *Advertising moving towards other communication platforms*

There are signs that also advertisers are leaving email and are moving towards new Internet communication platforms. While for malicious or threatening mail it is possible to have an idea of its ratio with respect to total email traffic by exploiting data reports from antispam filters, we have no data on the percentage of legitimate advertising mail. Despite this issue, under the simple hypothesis that legitimate advertising email is proportional to the budget allocated by companies for email marketing, we can derive other indexes of how the use of email as an advertising medium is changing in time, affecting the distribution of traffic generated for (legal) marketing purposes. In this way we see that the reduced growth in email traffic can also be attributed to another depletion phenomenon: given new "hot" media to spread their messages on, the legitimate advertisers decrease the budget quota dedicated to email marketing, resulting in proportionally less marketing emails being sent. With data derived by surveys on budget plans, [40] shows that in 2010 69% of interviewed companies increase the budget quota for Social Media, while 59% increase the budget for Email Marketing; in previous year surveys [41, 42] is reported that for Retail / Ecommerce industry fields the budget quota increased in 51% of cases for email, and in 79% for social media. This is reflected by the revenue from the different types of advertising, related to the relative effectiveness of email compared with other Internet marketing methods: from [43] in 2009 we see revenues from email decreasing of 28% of the value reported in 2008, becoming 1% of total Internet advertising revenues, while it accounted for 2% from 2004 to 2008, and 4% from 2001 to 2003 [44].

3.2 Looking at the future

Email once was the preferred medium for formal and work-related communication, providing identifiability (up to a degree), archiving, asynchronous communication (and also personal information management). However the speed of email exchanges (near-real-time) on the one hand, and the increasing pool of features offered by OSNs (broadcast-like with tweeter and status changes, archiving solutions, off-line retrieval), have shaded the differences, leading to similar usage patterns. Moreover, the increasing use of corporate OSNs profiles by enterprises "going 2.0", also with emerging adoption of intra-corporate OSNs (e.g. IBM beehive) are likely to have an impact on the use of email even for the professional world, as we have seen for advertising and marketing in general. OSNs adoption is showing a pattern strictly analogous to one seen before for email. People used to have multiple email accounts to keep work-related communications separate from personal ones (also allowing for continuity of personal bonds across job changes), or to provide different identities in different contexts (thematic mailing lists), in a similar way, due to the availability of different OSNs (characterized e.g. on the basis of language, or in being dedicated to a community of interest, or providing specific features and services), a single user can now use accounts on different OSNs according to the kind of social bonds he/she wants to manage, the intended audience for the specific message, and the type of interaction (cfr. poke, chat, wall post/blog entry, news sharing). As email did evolve features to provide an integrated management of multiple accounts (MUAs fetching from different accounts, switching among multiple *identities*; MTAs providing role-based aliases and automatic forwarding), the same evolution can be seen for OSNs: the wealth of different personal communication platforms gave birth to aggregating portals (like the ones that collect clickstream data used in [30]) and applications that offer a single homogeneous interface to multiple OSNs, IMs, VoIP calls, Video calls, **and email** (e.g. Empathy, Pidgin, Seesmic Desktop, Threadsy, Meebo, Google+). In such a unified communication environment, the technology behind the messaging platform is hidden, as the user receives and transmits messages seamlessly across multiple channels at the same time. One big architectural difference though is striking: email is a decentralized system, without a single controlling entity, based on international standards for interoperability, while on the other side the current OSNs are centrally administered and controlled; no democratic debate has been held to state their protocols, no peer opinion has been asked for the features to be included: each and any characteristic has been dictated by marketing reasons. The loss of control implied by the adoption of OSNs as widespread, general purpose communication medium, posed the well known risks to user privacy, and also on the availability of a tool that become central for personal and professional life of people (cfr. the possibility of forcibly closing of Facebook accounts and groups – allowed in the usage agreement accepted upon signing-in, and then lawful, but not fair to the users, left with no recourse to the law –, control on Facebook applications and features in general, – unidirectional – changing of usage policies). The possible relative death of email could be far from an innovation.

4. CONCLUSION

Since the raise of OSNs, people have had new means, alternative to email, to communicate personal messages. This has been provided without the need to install dedicated applications, without learning new usage routines, everything packed in the familiar *read, comment* loop made popular by WEB 2.0 interactive style. The new ingredient that was missing or too shattered in blogs, wikis and forums is the automatic integration of links to related people (involved in the discussion), related messages, and the like, which is natively provided by OSNs. In this complex online communication ecosystem, we analyzed email usage against alternative Internet-mediated communication tools, leveraging the

fact that the displacement of communication preferences of users affects the amount of traffic conveyed by SMTP and is thus measurable.

In order to analyze this evolution we performed an integrated ***multiple-viewpoints***, ***multi-variables*** and ***multi-sources*** analysis based on **20** different datasets, comprising network traffic traces, OSNs usage logs, Internet usage statistics, malicious or abusive email reports, and surveys about online advertising. The drawbacks and the advantages of the adoption of such sources of information have been detailed; the possible interactions among them and the related motivations are also provided.

Applying the proposed approach, we have shown how the comparison of traffic traces against Internet users and spam trends reveals a shift in the use of email using SMTP on TCP port 25, similar to what happened for the hard mail several years ago: being gradually relegated to specific - often formal - uses. A sign of this phenomenon can be seen in the change on the average length of email messages, that has become longer in time, because short, casual messages have been sent over other, more handy, applications. The steady fall of spam percentage in the last three quarters of 2008, can be interpreted also as the early symptom that spammers are leaving email for new Internet communication platforms (attracted to where the hype and better opportunities show up). In the future the overall trend depicted could become stronger due to the spreading use of novel integrated communication platforms, such as the communication environment offered by Google's portal, presenting email altogether with instant messaging, forums, audio and video calls, and social networking in a seamless interface. In this paper, we aimed at opening a discussion and a fresh debate on the topics here proposed; the availability of several other traffic traces (with different spatial and temporal features) will help in confirming, and going into more detail of, the depicted phenomena. Finally, the analysis and the methodology proposed in this paper could be also used to study other phenomena like censorship [53] and the impact on the network of natural disasters [54].

5. ACKNOWLEDGEMENT

We are grateful to the Editors and the anonymous Reviewers, whose comments helped us improving the quality and the content of the paper. The research activity described in this work has been partially funded by LINCE project of the FARO programme jointly financed by the Compagnia di San Paolo and by the Polo delle Scienze e delle Tecnologie of the University of Napoli "Federico II".

6. REFERENCES

[1] C. Partridge, "The Technical Development of Internet Email", Annals of the History of Computing, IEEE , vol.30, no.2, pp.3-29, April-June 2008

[2] R. Clayton, "Email Traffic: a quantitative snapshot", CEAS 2007, Mountain View CA, USA, Aug 2–3 2007

[3] R. Ohri, E. Chlebus, "Measurement Based E-mail Traffic Characterization", SPECTS'05, Philadelphia, PA, July 05.

[4] A. Dainotti, A. Pescapé, and G. Ventre, "A Packet-level Characterization of Network Traffic", CAMAD 2006, pp. 38–45, Trento (Italy), June 2006.

[5] A. Botta, A. Dainotti, A. Pescapé, G. Ventre, "Searching for Invariants in Network Games Traffic", ACM CoNEXT '06.

[6] L. Shyu, S. Y. Lau, P. Huang, "On the Search of Internet AS-level Topology Invariants", IEEE GLOBECOM 2006.

[7] S. Floyd, V. Paxson, Difficulties in simulating the Internet. IEEE/ACM Trans. Networking, Vol. 9 , Issue 4, pp. 392-403, Aug. 2001.

[8] M. J.-H. Lim, M. Negnevitsky, J. Hartnett, "E-mail Traffic Analysis Using Visualisation and Decision Trees", ISI 2006, vol. 3975 / 2006, pp. 680–681, San Diego, CA.

[9] F. Menges, B. Mishra, G. Narzisi, "Modeling and simulation of e-mail social networks: a new stochastic agent-based approach", 40th Conference on Winter Simulation (Miami, Florida, December 07 - 10, 2008), pp. 2792-2800.

[10] P. Svoboda, W. Jarner and M. Rupp, "Modeling e-mail traffic for 3G mobile networks", 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2007.

[11] C. Zhu, A. Kuh, "On Randomly Evolving Email Networks", Information Sciences and Systems, 2006 40th Annual Conference on , vol., no., pp.894-898, 22-24 March 2006.

[12] R.E. Filman, "When email was good", Internet Computing, IEEE , vol.7, no.3, pp. 4–6, May-June 2003.

[13] C.Y. Tseng, M.S. Chen, "Incremental SVM Model for Spam Detection on Dynamic Email Social Networks", CSE '09, vol.4, pp.128–135, 29-31 Aug. 2009.

[14] X. Sun, Q. Zhang, Z. Wang, "Using LPP and LS-SVM for spam filtering", CCCM 2009, vol.2, pp.451–454, 8-9 Aug. 2009.

[15] "Inside the Inbox: Trends for the Multichannel Marketer", *Epsilon's Global Consumer Email Study*, June 2009.

[16] `http://tracer.csl.sony.co.jp/mawi/`

[17] `http://www.wand.net.nz/wits/auck/8/`

[18] G. Dewaele, K. Fukuda, P. Borgnat, P. Abry, K. Cho, "Extracting hidden anomalies using sketch and non gaussian multiresolution statistical detection procedure." ACM SIGCOMM LSAD 2007, pp. 145-152 (2007).

[19] R. Fontugne, P. Borgnat, P. Abry, K. Fukuda, "Mawilab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking". ACM CoNEXT 2010, Philadelphia, PA, p. 12 (2010).

[20] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, K. Cho, "Seven years and one day: Sketching the evolution of internet traffic". IEEE INFCOM 2009, Rio de Janeiro, Brazil (2009).

[21] `http://www.internetworldstats.com/emarketing.htm`

[22] `http://www.maawg.org/about/EMR`

[23] `http://www.en.contactlab.com/`

[24] R. E. Grinter, L. Palen, "Instant messaging in teen life". ACM Conference on Computer Supported Cooperative Work (New Orleans, Louisiana, USA,

November 16 - 20, 2002). CSCW '02. ACM, New York, NY, 21-30.

[25] H. Smith, Y. Rogers, M. Brady, "Managing one's social network: Does age make a difference?," In Proceedings of INTERACT 2003, Zurich, 551-558.

[26] H. Ebel, M.-I. Mielsch, and S. Bornholdt, "Scale-Free Topology of E-mail Networks", Physical Rev. E, vol. 66, p. 035103(R), 2002.

[27] B. Viswanath, A. Mislove, M. Cha, K. P. Gummadi, "On the evolution of user interaction in Facebook", WOSN '09. ACM, New York, NY, 37-42.

[28] S. A. Golder, D. Wilkinson, and B. A. Huberman. "Rhythms of social interaction: Messaging within a massive online network". 3rd International Conference on Communities and Technologies (CT2007). East Lansing, MI., June 2007.

[29] F. Schneider, A. Feldmann, B. Krishnamurthy, W. Willinger, "Understanding online social network usage from a network perspective". 9th ACM SIGCOMM Conference on internet Measurement Conference (Chicago, Illinois, USA, November 04 - 06, 2009). IMC '09. ACM, New York, NY, 35-48.

[30] F. Benevenuto, T. Rodrigues, M. Cha, V. Almeida, "Characterizing user behavior in online social networks". 9th ACM SIGCOMM Conference on internet Measurement Conference (Chicago, Illinois, USA, November 04 - 06, 2009). IMC '09. ACM, New York, NY, 49-62.

[31] Sophos Plc.(2009) "Security threat report:2009", White Paper. https://secure.sophos.com/security/whitepapers/sophos-security-threat-report-jan-2009-na

[32] MessageLabs Inc.(2008) "MessageLabs Intelligence: 2008 Annual Security Report" http://www.messagelabs.co.uk/mlireport/MLIReport_Annual_2008_FINAL.pdf

[33] MessageLabs Inc.(2009) "MessageLabs Intelligence: 2009 Annual Security Report" http://www.messagelabs.com/mlireport/2009MLIAnnualReport_Final_PrintResolution.pdf

[34] Cisco Systems, Inc.(2009) "2009 Cisco Annual Security Report" http://www.cisco.com/en/US/prod/collateral/vpndevc/cisco_2009_asr.pdf

[35] Cisco Systems, Inc.(2010) "Cisco 2010 Midyear Security Report" http://www.cisco.com/en/US/prod/collateral/vpndevc/security_annual_report_mid2010.pdf

[36] G. Brown, T. Howe, M. Ihbe, A. Prakash, K. Borders, "Social networks and context-aware spam." CSCW '08, 2008 ACM conference on Computer supported cooperative work, pages 403412, New York, NY, USA, 2008. ACM.

[37] J. Baltazar, "Web 2.0 Botnet Evolution - KOOBFACE Revisited", white paper, TREND Micro Inc. http://us.trendmicro.com/imperia/md/content/us/trendwatch/researchandanalysis/web_2_0_botnet_evolution_-_koobface_revisited__may_2010_.pdf

[38] J. Baltazar, J. Costoya, R. Flores, "Show Me the Money - The Monetization of KOOBFACE", white paper, TREND Micro Inc. http://us.trendmicro.com/imperia/md/content/us/trendwatch/researchandanalysis/koobface_part3_showmethemoney.pdf

[39] Symantec Threats and Risks, Symantec Corp. http://www.symantec.com/security_response/writeup.jsp?docid=2008-080315-0217-99

[40] Marketing Sherpa LLC, 2011 B2B Marketing Benchmark Report http://ftp.marketingsherpa.com/Marketing\%20Files/PDF\%27s/Executive\%20Summary/2011B2B_BMR_ExecutiveSummary_100930.pdf

[41] Marketing Sherpa LLC, 2010 Email Marketing Benchmark Report http://www.marketingsherpa.com/EmailMarketingReport2010ESum.pdf

[42] Marketing Sherpa LLC, 2010 Social Media Marketing Benchmark Report http://www.marketingsherpa.com/SocialMediaExcerpt.pdf

[43] IAB Internet Advertising Revenue Report, 2009 Full-Year Results, April 2010 http://www.iab.net/media/file/IAB-Ad-Revenue-Full-Year-2009.pdf

[44] IAB Internet Advertising Revenue Report conducted by PricewaterhouseCoopers (PWC) http://www.iab.net/insights_research/1357

[45] The Nielsen Company, March 2009 "Global Faces and Networked Places." Report. http://blog.nielsen.com/nielsenwire/wp-content/uploads/2009/03/nielsen_globalfaces_mar09.pdf

[46] The Nielsen Company, News August 2010 "What Americans Do Online: Social Media And Games Dominate Activity" Report. http://blog.nielsen.com/nielsenwire/online_mobile/what-americans-do-online-social-media-and-games-dominate-activity/

[47] G. Vaudreuil, SMTP Service Extensions for Transmission of Large and Binary MIME Messages, RFC 3030 (Proposed Standard) (2000).

[48] http://www.grid.unina.it/Traffic/Traces/ttraces.php

[49] A. Dainotti, A. Pescapé, P. Salvo Rossi, F. Palmieri, G. Ventre, "Internet Traffic Modeling by means of Hidden Markov Models", Computer Networks (Elsevier), Volume 52, Issue 14, 9 October 2008, Pages 2645-2662.

[50] T. Phillips, A. Rosenfeld, A. C. Sher, "O(log n) bimodality analysis", Pattern Recognition, Volume 22, Issue 6, 1989, Pages 741-746, ISSN 0031-3203, 10.1016/0031-3203(89)90010-1.

[51] D.W.Scott, On optimal and data-based histograms, Biometrika 66, pp.605-610.

[52] S. Pederson and M. Johnson, Estimating Model Discrepancy, Technometrics, vol. 32, no. 3, Aug. 1990, pp. 30514.

[53] A. Dainotti, C. Squarcella, E. Aben, K. C. Claffy, M. Chiesa, M. Russo, A. Pescapé. "Analysis of country-wide internet outages caused by censorship", ACM SIGCOMM conference on Internet measurement conference (IMC '11). pp. 1-18, ACM, NY, USA.

[54] Z. S. Bischof, J. S. Otto, and F. E. Bustamante, "Distributed systems and natural disasters: BitTorrent as a global witness", Special Workshop on Internet and Disasters (SWID '11). ACM, NY, USA, Article 4, 8 pages.