

Worm Traffic Analysis and Characterization

Alberto Dainotti, Antonio Pescapé, and Giorgio Ventre
University of Napoli "Federico II" (Italy), {alberto,pescape,giorgio}@unina.it

Abstract—Internet worms are gaining ever more attention by the research community, representing one of the hot research topics in the field of network security. Our knowledge of phenomena related to Internet worms (from their intrinsic characteristics to their impact and to possible countermeasures) is still in its infancy. This is one of the main reasons for the existence of different kinds of research approaches. In this paper we focus on worm traffic analysis. We propose a general methodology, we discuss issues involved, and we present a software platform which can be used for this kind of study. Moreover, we show some interesting preliminary results from our traffic analysis of two of the most relevant worms that spread over the Internet: Witty and Slammer. Our results provide interesting evidences of (spatial and temporal) invariance and give some hints on worm traffic fingerprinting.

I. INTRODUCTION

Computer worms have come to the public attention as small software, usually written by a single individual, capable to take down the Internet [1]. A small, self-replicating, portion of code which can rapidly spread over hundred thousands systems, generating an overwhelming amount of overall traffic and consuming huge computational power. In 2001 the Code Red I and II worms spread all over the world by exploiting a bug in Microsoft web servers, causing denial of services, systems and network compromise, and links overload, corresponding to several billions damage [2]. In 2003, the Slammer worm, the fastest worm ever, spread to 90% of all potential targets in less than 10 minutes [3], reaching its full scanning rate - more than 55 million scans per second - in approximately 3 minutes. The Witty worm, the first to carry disruptive payload, spread in March 2004. Ironically, it infected hosts proactive in securing their networks [4]. Worms differ in scanning rate limitations (by latency or bandwidth), infection vector typology, scanning strategy, activity on the infected machine (damage, backdoor, attacks), etc. There is a rich literature of worm studies aiming at characterizing and modeling how the infections spread across networks, and of research works on worm detection and containment, based on many different approaches. On the other side, there is not much work related to a detailed analysis of the traffic generated by worms and comparing it to other traffic categories. This is testified not only by the lack of literature, but also by the lack of (i) methodologies and software tools for worm traffic analysis and (ii) traffic traces which can be considered useful for this purpose.

In this paper, we propose a first approach to the analysis and characterization of worm traffic; discussing the difficulties

⁰This work has been partially supported by the MIUR in the context of PRIN 2006 RECIPE Project, by CONTENT EU NoE, OneLab, and NETQOS EU projects. The authors would like to thank Alessandro De Peppo for his incredible work on data analysis.

currently present, and how they can be overcome. We present our software platform to extract new data, sanitize traces, and analyze traffic properties. Finally we show preliminary results of this analysis made on Witty and Slammer worms, studying and comparing traffic from three network links. Results show some interesting properties (among them, time and space invariance) and some peculiarities of worm traffic which make it different from other categories of traffic commonly found on the Internet. Besides representing the first steps into more deeply understanding worm traffic and how it can affect links, such results can be also considered for the design of new fingerprinting and detection techniques.

II. MOTIVATION AND RELATED WORKS

Recently, computer worms have been subject to several studies, and research efforts are made in different directions. To better understand related literature and how our work fits in, we can identify the following main areas: worm behavioral characterization, spread modeling, detection, traffic characterization.

As for the first point, a comprehensive classification of computer worms is presented in [5]. Moreover, an analysis of specific worms (Witty [4], Slammer [6] [3], Code-Red [2]) is presented in several works. However, the results related to their traffic are basic: how and at which speed the worm spread, the scanning strategy and rate achieved, the distribution of IPs contacted, and how the packets are built. Sometimes the aggregate packet rate of worm traffic on a link is shown. Such information constitutes a valuable analysis of the worm characteristics, which is a fundamental first step to understand worms, how they work and their impact, and to build deeper works on top of that. Another research field involves modeling the spread of worms using analytical and simulative approaches, also taking into account the effects of patching, human countermeasures, and congestion caused by the worms themselves (e.g. [7] [8]). Such studies can be used to design worm containment strategies. As regards detection techniques, they can be differentiated mainly in two approaches: content based and traffic based. Content inspection approaches can be based on signatures of known worms or on correlation of common patterns found in packets to detect new worms during their initial spread [9]. Content analysis, however, requires heavy resource consumptions and can be made ineffective by mutant worms. In contrast, detection methodologies based on traffic observations are related to the probing behavior of scanning worms. A common approach is to identify illegitimate scans and the increase of activity due to worm propagation. In [10] hypothesis testing is used to detect infected hosts

by monitoring the number of failed connections they try to initiate. A similar approach was previously proposed in [11]. Relying on failed connections is obviously not viable for worms based on single UDP packets (e.g. Witty and Slammer). In [12] infected hosts are detected by intercepting packets they send to inactive addresses. Then, a change point detection technique is applied to the rate of infected hosts. The works [13] [14] can be partially ascribed both to the research areas of detection and of worm traffic characterization. Because, starting from the observation of statistical properties of worm traffic - the exponential growth trend of infections at the early propagation stage - they propose two detection techniques.

As regards the area of traffic characterization, there are not much more works related to worms. Indeed, while in the past years several insights on statistical properties of aggregate and specific application (Web, network games, file transfers, multimedia, ..) traffic have been gained, not much is known about *unwanted* traffic, and worms in particular. However, it has been demonstrated that understanding the statistical properties of traffic at different levels (aggregate, flows, sessions, packets) can bring important results. In [15] an active approach to understand some properties of all unsolicited traffic is adopted. However there is no specific characterization of worm traffic. Whereas in [16] and [17], other kinds of anomalous traffic - not worms - have been studied: Distributed Denial of Service and Flashcrowds. The multi-resolution analysis of their traffic shows that flash-crowds and DDoS have different properties in terms of marginal distributions and of covariance. They show that the properties found can affect link QoS, and apply the analysis results for detection purposes.

The approach presented in this paper is a first attempt to fill the lack of literature, complementing the other works by focusing strictly on the characterization of worm traffic. This means to analyze properties of marginal distributions, of time dependence, of time-scale analysis etc. looking at traffic at different levels. The characterization is done by comparing results from different data sets, taken at different sites and at different times, and possibly related to more worms. Moreover, a comparison is made between worm and legitimate traffic, and the impact of worms on traffic found in links under normal conditions is investigated. This kind of approach permits to better understand how the presence of worm traffic impacts on network links and nodes, and the gained knowledge could be exploited for fingerprinting and detection purposes.

III. APPROACH

The general approach proposed in this work consists in analyzing statistical properties of traffic generated by worms, looking at it from different points of views. That is, analyzing properties of aggregate traffic, but also separating it into streams - we will call them *sessions* - by source hosts, or by flows, etc. and considering not only sessions-related variables (as arrival times, size, duration, etc.) but also packet-level variables inside sessions: inter-packet times (IPT) and packet sizes (PS). We are also interested in comparing findings with other categories of traffic (i.e. non-worms). The

proposed passive approach is based on the observation of traffic both traversing backbone links and captured by network telescopes (i.e. directed towards unused addresses). Because of the novelty of this work, we also discuss issues involved in performing such kind of analysis (see Section IV-D), as the lack of useful traffic traces and the need for data sanitization. Moreover, we present a software platform useful at various stages of such kind of study.

To give a clearer idea, the general approach used can be synthetically sketched into a number of sequential steps (with possible feedback lines) depicted in Fig. 1. After the traffic



Fig. 1. Life Cycle of Data Analysis.

trace acquisition (first block), human intervention is usually necessary to inspect the trace. Understanding which kind of traffic has been captured is a first fundamental step before performing a detailed statistical analysis. To do this, we need flexible tools to rapidly investigate several properties of traffic, from looking into headers and payload (if present) to reporting concise information on hosts, flows, etc. From this analysis it is possible to choose on which aspect we want to focus the characterization and to conceive strategies for automated trace sanitization to remove spurious data (step 3). For example, as Witty and Slammer worms send a single UDP packet to each victim host: (i) by analyzing reports related to flows, we could immediately spot non-worm behaviors looking for flows with more than one packet; (ii) we learnt that, after isolating worm traffic, flow-level analysis was not of interest to us. In the next step, the software tool extracts measurements data from the traffic trace and it may also be able to perform a preliminary analysis. For example, we used threshold mechanisms, based on packets parameters, to automatically mark hosts as infected on the basis of some typical symptoms. Finally, the data sets obtained can be loaded into statistical analysis software and analyzed, looking at marginal distributions, time dependence, correlations etc. Moreover, frequency/scale and time localization analysis can be performed by means of the Fourier and Wavelet transforms.

When analyzing the data, we look for repeating behaviors (the “*search for invariants*” [18]) and, by applying the same analysis to worm and legitimate applications, we aim at sketching similarities and differences. Also comparing the overall traffic before and during worm propagation may allow to infer information about the impact of worms on links and nodes. As anticipated, in this paper we present a preliminary analysis of two worms: Witty and Slammer. The results here shown are basically related to aggregate traffic and to the analysis of host-based sessions, focusing on packet-level variables: IPT and PS (for details see Sec. IV-C). We apply a packet-level analysis already adopted for the traffic generated by

legitimate applications [19] [20]. Packet-level analysis offers indeed several advantages. One of the most important is that, being independent of the application-level protocol, it can be equally applied to different kinds of traffic. Furthermore, characterizing statistical properties of traffic at packet-level can help in building analytical and empirical models to be used for traffic generation and simulation, which represent another mean to better assess the impact of worm traffic on links and nodes. Finally, traffic at packet level remains observable after encryption made by, for example, end-to-end cryptographic protocols such as SSL or IPSec, making packet-level traffic modeling a robust approach to traffic profiling for anomaly detection and traffic classification.

IV. TRAFFIC TRACES AND TOOLS

A. Considered worms: Witty and Slammer

Because of the poor availability of worm traffic traces useful for traffic characterization (see Section IV-D), we limited our study to the Witty and Slammer worms, which will be briefly described here. Details on the traces and tools will be given in the next subsections. The Witty worm [4] exploited a bug in the ISS firewall software when decoding ICQ servers packets [4]. It sends a single UDP packet with source port 4000 to each scanned host. The payload varies from 768 to 1279 bytes, because of a random padding which is done to make worm identification (e.g. by firewalls) harder. After 20000 packets have been sent to randomly chosen IP addresses, it overwrites a small portion of the hard disk, and then it starts to send packets again. The Slammer worm [6] [3] instead, exploits a bug in Microsoft SQL Server. It sends a single UDP packet of fixed size (404 bytes) with destination port 1413 to each target. The scanning strategy is random. However, a bug in its random number generator left a considerable portion of the Internet hosts not scanned. Differently from other worms (that are latency-limited because they issue a *connect()* call for each host to be scanned), as for example Code Red, both Witty and Slammer are bandwidth-limited worms. This is because they send UDP packets and do not need to wait for any response from the potential victim. So they are only limited by the bandwidth of the infected machines.

B. Traffic Traces

In Tab. I the traffic traces that have been used in this work are summarized. As for the Witty worm, we analyzed several tens of gigabytes of data collected and made available by CAIDA [21]. The traffic stored in such files has been collected by a network telescope, that is, all the traffic directed towards an unused address space has been captured. This way, unsolicited traffic (e.g. automated scans) can be detected and observed. The traces here used have been obtained by filtering the traffic captured by the network telescope, in the days of the spread of the Witty worm, considering only UDP packets with source port 4000. Moreover, to obtain more traces related to Witty and to overcome the poor availability of worm traces, we looked into traffic traces of a trans-oceanic link during the days of the worm spreading, verifying the presence of

packets which can be associated to the Witty worm (second row of Tab. I). Indeed, the MAWI-WIDE project [22] makes available 15 minutes traffic traces of this link for each day of the year since 2000. An important benefit of such approach is that, in this way, we also have the availability of data related to legitimate traffic captured from the same link, and at the same time, of the worm related traffic. This is good to compare their properties. As explained in Section V-A, many results from the analysis show that the Witty traffic selected from this trace has consistent properties with that in the trace made available by CAIDA (evidence of spatial invariance).

We also looked into MAWI traffic traces captured during the spread of Slammer. But we could not find packets associated to this worm. This is probably due to a filtering rule which was set on the routers. Traffic traces related to Slammer have been made available by MIT [6]. They were obtained by filtering all the traffic traversing two unidirectional links, considering only UDP packets with port 1413. These traces have been collected on March 25th, 2004, which was one of the days in which Slammer activity was highest. All the traces in this work contain only packet headers until layer 4, that is, no payload information is stored.

C. Tools for data capture and analysis

For the activities in the blocks 2-4 in Fig. 1, we extended our software platform, called Plab [23]. Plab is an open-source software we developed for analysis of live traffic and traces in tcpdump format. It was employed in previous works on traffic analysis and modeling [19] [20], but the features introduced in the latest release have been specifically designed for this work. Because of space constraints they can not be exhaustively presented here, giving only a brief overview we refer the reader to the software documentation for more information [23]. Plab is capable to efficiently analyze very large traffic traces and to separate traffic into different *sessions*. Depending on user-specified parameters, a session is identified by: (i) all packets sent and received by a host (*host mode*); (ii) all packets identified by source and destination IP and ports with a default timeout of 60 seconds (*flow mode*); (iii) all packets exchanged by 2 hosts related to a specific service (e.g. TCP port 80), with a user definable timeout (*conversation mode*). Given one of the above modes, sessions are assigned an ID, and for each session the IPT between packets flowing in the same direction are calculated, along with PS. We call such data *packet-level* data series. Moreover, the arrival time of each session, its duration, and bytes transmitted for each direction are calculated, allowing to perform an analysis at a higher level (host/flow/conversation level). IPT and PS looking at the traffic as a whole are also calculated. In this work we use host-based sessions. We added also specific features which were used for data sanitization (see Section IV-D). Data sets extracted by Plab are then processed under the Matlab environment. We developed a library of scripts, available at [23], which can be used for statistical analysis of traffic data, together with other tools made available by the research community (e.g. Wavelet analysis [24]).

TABLE I

TRACES DETAILS.								
Worm	Source	Observation point	Filter	Date	Duration	Size	Estimated Infected Hosts	
Witty	CAIDA	Net Telescope	udp src port 4000	March 20-22, 2004	15m per day	1.3GB	10725	
Witty	MAWI	BIDIR Link	ALL	March 20-22, 2004	15m per day	2.1 GB	4728	
Slammer	MIT	UNIDIR Link 0	udp dst port 1434	March 25, 2003	8h 44m	842 MB	2523	
Slammer	MIT	UNIDIR Link 1	udp dst port 1434	March 25, 2003	8h 44m	431 MB	5321	

D. On the data available to researchers

The approach we propose is about systematic characterization of worm traffic. In this section we report on the issues related to the poor availability of appropriate measurement data which researchers must face to carry out such study.

First of all, we register the scarce availability of worm traces in general. Moreover, even traffic traces used in research papers (e.g. Slammer [3] and Code-red [2]) are sometimes not made public. Another aspect is related to the characteristics of the available traces: when too small or sampled, they can be inappropriate to perform a careful traffic analysis/characterization. For example, the National Laboratory for Applied Network Research (NLANR) collects each day 8 traces of 90 seconds each from several backbone links in the USA. Among them, there are traces captured during the days of worms spread (of Code-red I and II, Slammer, MyDoom). However, traces of such a small length cannot be used to characterize sessions; also they do not allow to perform time/scale analysis on large time scales. In contrast, if larger they would be of great value for our purposes. Other backbone traces report only flow-level data: timestamp of flow start/end, packets and bytes transmitted, etc. Such kind of information does not allow to perform packet-level analysis. Furthermore, most of the available traces, as for example those from CAIDA and MIT used in this work, do not contain the rest of the legitimate traffic flowing on the links. This is because the observation point is a network telescope or because traces were deliberately filtered before making them available. This does not allow to do a fair comparison between the worm traffic and the legitimate traffic flowing on the same link at the same time. Also, it does not allow to study the effect of worm traffic on the overall aggregate traffic.

Finally, to obtain reliable results, traces need sanitization before analysis. Indeed, when traces are reported as containing only worm traffic they are usually filtered by port numbers or other simple indicators. Thus, it may happen to be also non-worm traffic inside the trace. Sometimes captured packet traces can contain spurious data due to hardware and software errors during data acquisition (as replication of data). Both legitimate traffic and spurious data were found in the MIT traces used in this work, and we added fingerprinting functionalities in Plab to aid the operator into identifying them. For example, we found in the *Link 0* MIT trace 13 hosts that generated only legitimate traffic (probably DNS). We found similar results for the *Link 1* MIT trace. Even if this traffic represents only a small fraction of the trace, finding such legitimate hosts and flows is important (i) to make reliable worm traffic characterization, and (ii) to compare worm traffic against legitimate traffic flowing at the same time and on the same link.

The scarce availability of usable data lead us into searching

for more traces. A minor contribution of this work is that we found Witty traffic (along with common traffic) in the traces captured from a MAWI backbone link (as explained in Section IV-B). We used the fingerprinting techniques implemented in Plab to complement and verify our filtering of Witty traffic from the MAWI traces which was based on selecting UDP packets originating from port 4000 and with PS > 768 bytes.

V. DATA ANALYSIS

A. Analysis of Witty Traffic

As for Witty, for each considered trace we built the list of the hosts generating most of the traffic. By analyzing a 15 minutes trace of 20 March 2004 from CAIDA, we found that on a total number of 7515 scanning hosts, the 26 hosts sending more than 30000 packets are accountable for only the 9% of the total trace traffic. For the corresponding MAWI trace we found that the 10 top hosts, on a total of 2881, are accountable for 9% of the total traffic. First of all, it is worth to note that we found two similar results from two different kinds of observation points: a transoceanic link and a network telescope (this represents an early evidence of *spatial invariance* and it paves the way to further investigations). Moreover, the corresponding analysis made for the Slammer traces revealed very different outcomes (see Section V-B). This is interesting because Witty and Slammer behaviors show strong similarities [25], as for example the scanning strategy and the sending of one single packet per victim.

1) *Marginal distributions of IPT and PS*: as stated earlier, we consider IPT inside host-based sessions (i.e. an IPT represents the time between two consecutive packets sent by the same host). We measure IPTs with a resolution of $1\mu s$ and apply a logarithmic transformation because they range over several orders of magnitude. For each trace the distribution of IPTs is built by putting together all the IPTs calculated for each host-based session. In the following, when necessary, we will refer to the corresponding PDF as the *average* PDF, to distinguish it from the PDF made by IPTs of a single session.

In Fig. 2 the diagrams of the IPTs CDFs and PDFs are depicted. The distributions are quite regular, resembling a gaussian distribution. This behavior is invariant with respect to the site observed and to the time (again some properties of *invariance* in both time and space). As for the first point, the mean of the distribution is shifted when the link changes. This can be probably connected to the number of IPs which can be observed: the lower the number of victims per host, the larger the average IPT. As for the *invariance* with respect to time, we observe that the IPT distribution derived from a specific observation point does not change in the different epidemical stages. Indeed, as can be seen from Tab. II, while the first day represents the explosion of the epidemic, in the

subsequent days the infection level decreases dramatically, probably because of patching (see infection models taking patching into account [8]). Also, Tab. III shows that, for each site, all the distribution statistics but the entropy¹ keep approximately the same values as the considered day changes. In contrast, the entropy follows a descending trend as the infection decreases, possibly because the number of infected hosts decreases thus reducing the *uncertainty* associated to the PDF. This finding suggests a possible application to identify the evolution status of a worm spreading.

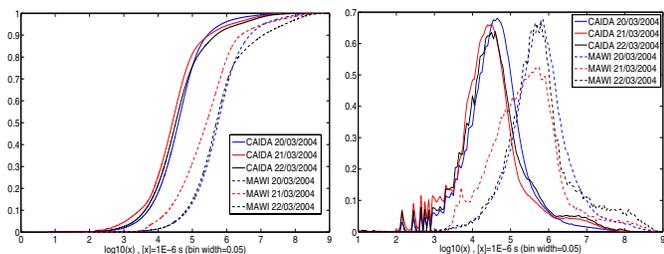


Fig. 2. Witty Inter-packet times.

TABLE II
WITTY TRAFFIC STATISTICS.

	CAIDA 20/3	CAIDA 21/3	CAIDA 22/3	MAWI 20/3	MAWI 21/3	MAWI 22/3
Pkts	9,261,414	2,986,325	701,314	226,034	102,727	33,941
Src Hosts	7,515	2,128	1,085	2,881	1,141	706
Dst Hosts	6,800,779	2,690,668	683,096	198,663	99,380	33,231

TABLE III
WITTY IPT ($\log_{10}(x)$, $[x] = 1E - 6s$).

Trace	Mean	Median	Max	StdDev	Entropy (bit)
CAIDA 20/3	4,593	4,459	8,870	0,750	7,825
CAIDA 21/3	4,454	4,415	8,845	0,873	7,247
CAIDA 22/3	4,592	4,513	8,717	0,885	6,570
MAWI 20/3	5,762	5,750	8,898	0,720	6,112
MAWI 21/3	5,413	5,410	8,904	0,902	5,710
MAWI 22/3	5,802	5,670	8,921	0,902	5,118

Another interesting aspect that came out from the study of Witty traffic is related to the payload size. As anticipated, this worm is designed to pad the packet payload with a random number of bytes. In Fig. 3 the CDF and PDF diagrams of the PS of packets sent by Witty hosts from both CAIDA and MAWI traces of three different days are depicted and compared. The figures show that in all cases the distributions can be well approximated by a uniform distribution from 768 to 1279 bytes, which is totally different from typical payload size distributions commonly found on Internet links [26].

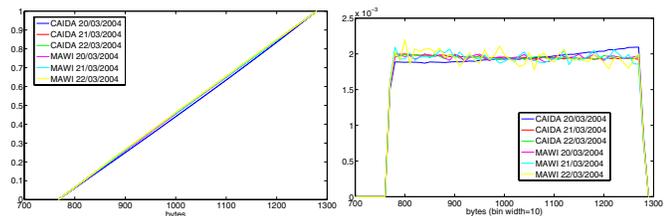


Fig. 3. Witty Payload Size.

A bivariate PDF diagram puts together information related to the PS and IPT marginal distributions, taking into account

¹The entropy is calculated as $-\sum_i P(x_i) \cdot \log_2 P(x_i)$ where $P(x_i)$ is the probability associated to each bin (of width 0.05) of the samples histogram.

also mutual dependencies between the two variables. In Fig. 4 the bivariate PDF diagrams of Witty traffic, related to MAWI and CAIDA are shown. They are very similar, confirming a typical behavior, from a traffic characterization point of view, of the infected hosts. This is an interesting invariant, which makes such diagrams (or the information contained) to be considered for fingerprinting and detection techniques.

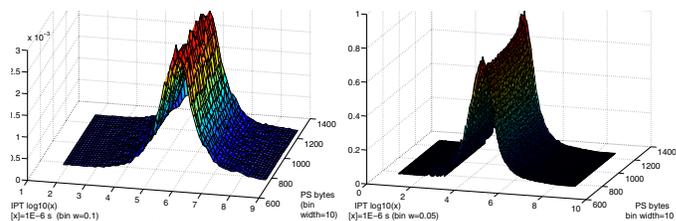


Fig. 4. MAWI(left) and CAIDA(right) Bivariate PDF (20/03/2004).

To understand how much the traffic properties we found are really peculiar to Witty, we made some comparisons against legitimate traffic. In this process, in order to make comparisons more meaningful, we tried to remove all possible differences due to side-effects. For this reason, we chose DNS traffic as an example, because it runs on the same transport protocol - UDP - of Witty (in contrast, TCP end-to-end flow control could somehow affect packet-level variables) and a DNS server, like a worm-infected host, talks to several different hosts in a short time. Moreover, the DNS traffic analyzed is from the same MAWI trace of Witty, therefore there are no link-dependent or time-dependent aspects which could be differently influenced. In Fig. 5, the bivariate PDF of PS and IPTs calculated for the 4593 DNS servers (hosts sending packets only from source port UDP 53) found in the MAWI trace of 20th of March shows a totally different profile. The DNS packet payloads are rarely larger than 250 bytes, with an heavier concentration around three byte-lengths, and IPTs are spread but with the main peaks in the first decade and in the region between the fourth and seventh. To stress the concept of the possible

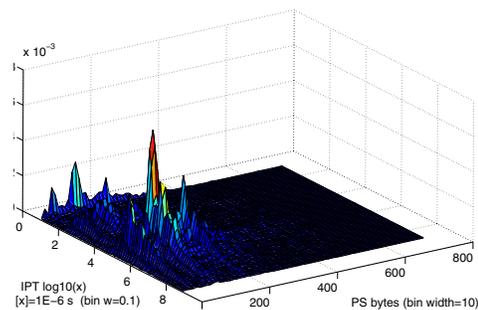


Fig. 5. MAWI DNS bivariate PDF.

application of such findings in the field of fingerprinting and classification, we show in Fig. 6 the bivariate PDFs obtained for two single hosts. On the left, one of the Witty-infected hosts, chosen among those generating more traffic, is shown. Whereas the diagram on the right is related to the most active DNS server. The choice of a larger binning and the

less smoothness of surfaces are due to a reduced number of samples compared to the PDFs obtained by averaging data from all hosts. However we can see that: (i) the average bivariate PDFs reflect well the properties of the single hosts, and (ii) the bivariate PDFs of the two considered hosts are totally different.

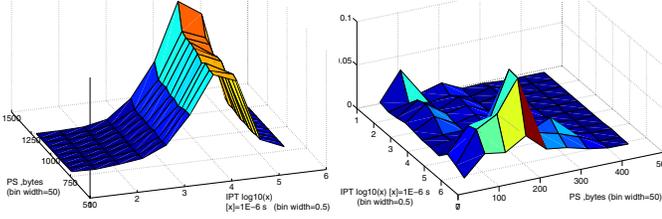


Fig. 6. Bivariate PDFs of single hosts: Witty (left) and DNS (right).

2) *Autocorrelation*: besides looking at marginal distributions, we studied PS from a time dependence perspective. For several hosts infected by Witty, we plotted the sequence of PS (Fig. 7) and its autocorrelation function from lag 0 to 100 (Fig. 8), and compared them to the corresponding ones generated by DNS servers found in the MAWI trace. Both kinds of graphs can clearly highlight the different behavior of a *Witty-infected* host from a legitimate host (a DNS server). The sequence of Witty PS is totally uncorrelated, whereas there are strong indications of correlation in DNS traffic. The uncorrelation

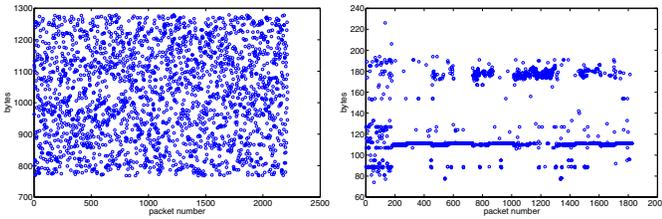


Fig. 7. PS Sequence Graph: a Witty host (left) and a DNS server (right).

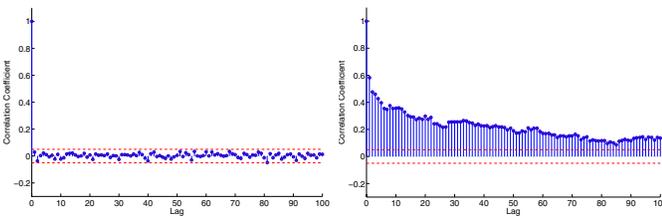


Fig. 8. PS Autocorrelation: a Witty host (left) and a DNS server (right).

of Witty payloads is obviously a consequence of the random padding, whereas the presence of correlation in DNS traffic might be explained by the application-protocol structure and by the content of the DNS reply. It is also interesting, however, that we found a similar distinction as regards packets IPTs. Those observed from Witty hosts are uncorrelated at all lags (both for CAIDA and MAWI traces), whereas for DNS hosts IPTs are correlated at several lags (Fig. 9).

3) *Wavelet Analysis*: in this sub-section we show results of a wavelet-based analysis aimed at understanding if Witty traffic affects aggregate traffic temporal structures from a multi-resolution point of view. Studying this aspect is important because it has been shown in literature that scaling properties like Long Range Dependence (LRD) are frequent in network traffic and they can have a negative impact on network performance. We adopt the estimation technique exposed in [24], based on the *Discrete Wavelet Transform* of a random process X of size N . A dyadic decomposition is applied, so that the number of considered scales is $J \approx \log_2(N)$. The so-called *Logscale Diagram (LD)* shows the trend followed by (the logarithm of) the energy of the wavelet coefficients at each scale, allowing to estimate the scaling behavior of the process X and the *Hurst* parameter (which is an indicator of LRD and self-similarity). We considered the time series of packet rate sampled with a period of $1ms$. That is, the i -th sample corresponds to the number of packets sent between the i -th and the $(i+1)$ -th milliseconds from the start of the trace. To make comparisons of the Wavelet spectrum easier, the packet rates of the whole link traffic, or related only to Witty, have been normalized so that their energy was equal to 1. For each time series X , we divided each sample by $X_{rms} = \sqrt{\sum_{x_i \in X} (x_i^2)}$.

In the left diagram of Fig. 10 we show the LD of MAWI traffic from the days 6/3 13/3 20/3 and 3/4 (all Saturdays). It is easy to notice that while the other LDs approximately follow the same trend, the LD corresponding to 20/3 (the day in which the Witty worm was spreading) departs from the other ones from scale 15 to scale 17. The right diagram of Fig. 10 shows the LD of the only Witty traffic extracted from the trace of day 20/3. It is interesting to note that the trend is flat until scale 15. Thus, such results suggest that there could be an influence of worm traffic on the scaling behavior of the aggregate traffic on the link.

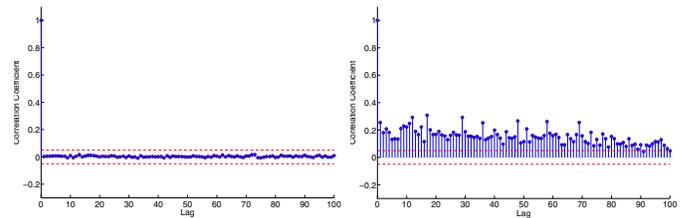


Fig. 9. IPTs Autocorrelation: a Witty host (left) and a DNS server (right).

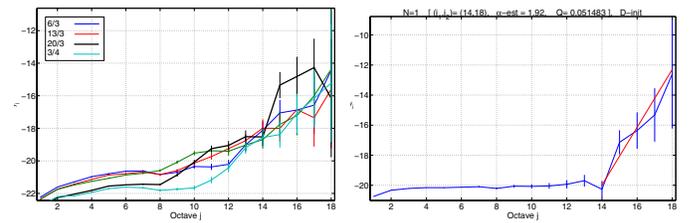


Fig. 10. Wavelet Analysis: MAWI total traffic (left), MAWI Witty (right).

B. Analysis of Slammer Traffic

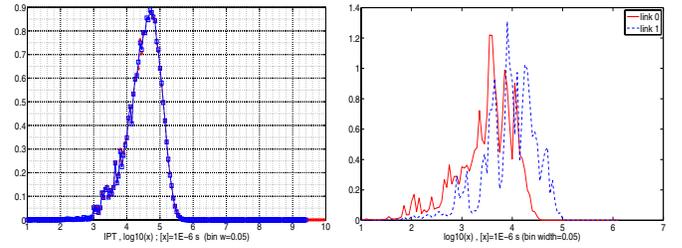
In Slammer MIT traces from *Link 0*, the 8 most active hosts are responsible for about 80% of all the traffic (on a total of 2523), whereas on *Link 1* 12 hosts (on a total of 5321) generate 75% of the whole. As said before, this behavior is different from what we found in the Witty traces captured from both MAWI and CAIDA. It seems that the rate at which the infected hosts send packets can vary largely, and that there are few hosts able to achieve much higher rates than all the others. Obviously, this can be due to the specific conditions of their access links and hardware involved, but it could be also influenced by the random scanning strategy adopted in Slammer, which makes some hosts to select more victims whose routes traverse the observed link. At this stage we limit to note this difference with Witty, in which we found more homogeneity among hosts also in other aspects of the analysis. Moreover, we observed that some Slammer-infected hosts had sudden changes in the average scan rate during the 8 hours traces. We will return on this subject later.

The Slammer traces availability does not allow to compare worm traffic of two totally different links/observation sites, neither offers the possibility to observe and compare it against large quantities of legitimate traffic flowing in the same links. However, they offer another interesting insight: the 2 MIT links belong to the same backbone and the anonymization algorithm applied to both traffic traces is the same. Therefore we were able to recognize the presence of some infected hosts in both traces, allowing to compare their behavior in both links.

1) *Marginal Distributions of PS and IPT*: as reported earlier, PS is fixed in Slammer. Therefore, from a packet-level point of view, we limited to the analysis of IPTs. Even from this aspect, Slammer hosts behave more heterogeneously than the Witty ones. The IPT PDFs calculated for the single hosts show more differences, however the average PDF is able to represent them. In all cases the profiles found were quite different from those of DNS traffic shown earlier, and from those of the small portion of legitimate traffic found on the original MIT traces. Moreover, thanks to the subnet membership preserving property of the anonymization algorithm used in the MIT traces, we were able to find in the same trace two infected hosts coming from the same network (moreover they belong to the list of the 12 top hosts). Fig. 11(a) shows that their IPT PDF profiles are very similar. This seems to confirm that parameters related to the access link (location, bandwidth, latency, etc.) have a significant influence.

In Fig. 12 the average PDFs and CDFs of Slammer hosts on both links are compared. The very similar shape of the curves is an interesting invariant. Whereas, the shift of the curves of *Link 1* when compared to those of *Link 0* is explained by the fact that the first link routes less traffic. This difference is testified by the mean and median values reported in Tab. IV, while the other parameters (e.g. entropy) show a general accordance.

As regards comparing results from the two links, we also observed the behavior of the same infected hosts present in



(a) IPT PDFs: 2 hosts from the same subnet. (b) IPT PDFs: the same host on 2 links.

Fig. 11. Slammer IPT.

TABLE IV

SLAMMER IPT ($\log_{10}(x)$, $[x] = 1E - 6s$).

Source	Mean	Median	Max	StdDev	Entropy (bit)
Link 0	4,046	3,918	10,47	0,953	8,00
Link 1	4,710	4,455	10,46	1,123	7,70

both traces. In general we noticed a very close behavior. An example, related to one of the hosts generating most traffic in both links, is shown in Fig. 11(b).

2) *Autocorrelation*: like the results found for Witty (and differently from DNS traffic), IPTs are basically uncorrelated for all the Slammer-infected hosts we analyzed. However most of them present a very small correlation which oscillates around 0 from lag 0 to lag 100. This can be observed from Fig. 13, where two different hosts from the two links are analyzed.

An interesting aspect came out during autocorrelation analysis. From the observation of the MIT traces, the infected hosts keep approximately the same packet rate all the time (when active). However, some of them suddenly change the packet rate to a different average for periods which can last from several minutes to hours, which can be explained e.g. with a change in the amount of available bandwidth on their access links. An example is reported in the left diagram of Fig. 14 for an host on *Link 0*. We observed that the analysis of the autocorrelation of IPTs is heavily affected by such behavior, bringing to large autocorrelation values. This can be observed by comparing the autocorrelation plot obtained

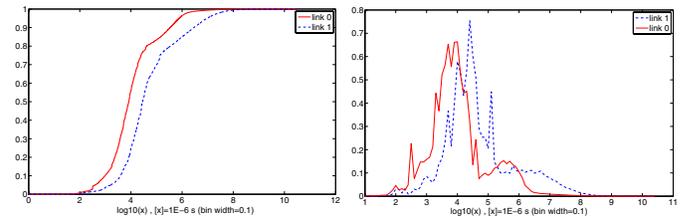


Fig. 12. CDFs (left) and PDFs (right) of Slammer IPTs.

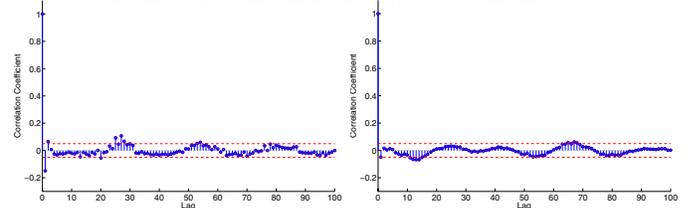


Fig. 13. IPT autocorr. of two Slammer hosts, on *Link 0* (left) and *Link 1* (right).

for the complete trace of the considered host (right diagram in Fig. 14) with the one obtained from the analysis of only the first 3 hours of data (left diagram in Fig. 13). Therefore, this must be taken into account when characterizing traffic of an infected host, and specifically if such results need to be used in the context of traffic classification and anomaly detection. On the other side, PDFs are not significantly affected by such behavior, probably because of the transformation of the IPT samples in the logarithmic domain.

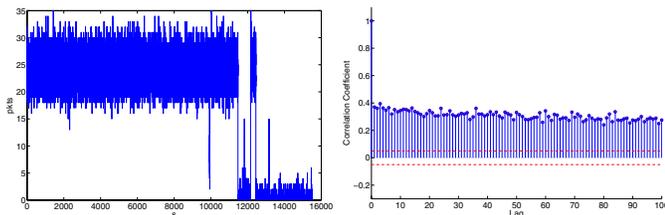


Fig. 14. A Slammer host: packet rate (left), IPT autocorrelation (right).

3) *Wavelet analysis*: as regards, multiscaling properties of Slammer traffic, we could not study how it affect overall link traffic because of the lack of data in the trace. However, from the LD applied to Slammer-only traffic shown in Fig. 15 we can notice that: the LDs of both links are very close. Moreover their trend is similar to the one of Witty traffic: a flat trend until a scale from which a clear scaling behavior starts. Also, the estimate of the slope in that region, given by the parameter α which determines the estimation of the Hurst parameter, is almost the same (about 2) for both links and for witty traffic.

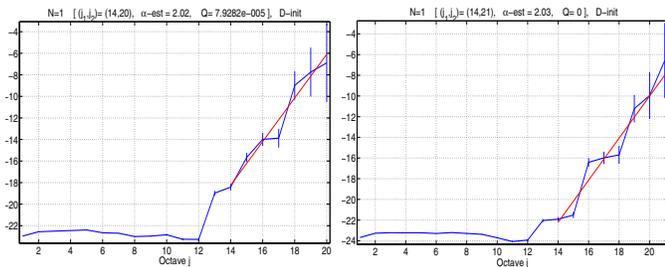


Fig. 15. Slammer Wavelet Analysis: *Link 0* (left) and *Link 1* (right).

VI. CONCLUSIONS AND FUTURE RESEARCH

The contribution of this paper is twofold. First, we propose an approach to the study of worm traffic, discussing a general methodology and current issues, and also presenting a software tool which can be used for this research. Second, we show preliminary results from the application of this methodology to two well-known worms. Such findings have shown that worm traffic presents interesting properties of (spatial and temporal) invariance, and looks very different from other kinds of traffic. This aspect is reflected by: (i) joint characterization of the marginal distributions of PS and IPT shown by means of bivariate PDFs; and (ii) lack of temporal (auto-)correlation in the PS and IPT time series. Moreover, some preliminary results on the entropy of packet-level statistics have been

found, and the Wavelet analysis seems to show that worm traffic has an impact on the statistical properties of aggregate traffic at multiple time scales. Therefore, this study represents a first step to understand more deeply the impact of worms on network traffic, giving insights which we plan to use for classification and detection purposes.

REFERENCES

- [1] S. Staniford, V. Paxson, and N. Weaver "How to Own the Internet in Your Spare Time" *11th USENIX Security Symposium*, 2002.
- [2] C. Shannon, D. Moore "The Spread of the Code-Red Worm". http://www.caida.org/analysis/security/code-red/coderedv2_analysis.xml
- [3] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, Nicholas Weaver "Inside the Slammer Worm" *IEEE Computer*, 2003
- [4] C. Shannon, D. Moore "The Spread of the Witty Worm". *IEEE Security & Privacy*, vol. 2, no. 4, pp. 46-50, Jul. 2004
- [5] N. Weaver, V. Paxson, S. Staniford, R. Cunningham "A Taxonomy of Computer Worms" *ACM CCS First Workshop on Rapid Malcode (WORM)*, Oct. 2003
- [6] Robert Beverly, MIT LCS "MS-SQL Slammer/Sapphire Traffic Analysis", 2003 <http://momo.lcs.mit.edu/slammer/>
- [7] C. C. Zou, W. Gong, D. Towsley, "Code Red Worm Propagation Modeling and Analysis", *9th ACM CCCS*, Nov 2002.
- [8] Z. Chen, L. Gao, K. Kwiat, "Modeling the spread of active worms", *IEEE INFOCOM 2003*
- [9] P. Akritidis, K. Anagnostakis, E.P. Markatos, "Efficient content-based detection of zero-day worms", *ICC 2005*, May 2005
- [10] S. Schechter, J. Jung, and A. W. Berger. "Fast Detection of Scanning Worm Infections", *7th RAID*, Sep. 2004
- [11] N. Weaver, S. Staniford, and V. Paxson, "Very fast containment of scanning worms", *USENIX Security Symposium*, Aug. 2004.
- [12] J. Xia, S. Vangala, J. Wu, L. Gao, K. Kwiat, "Effective Worm Detection for Various Scan Techniques", to appear in *Journal of Computer Security*
- [13] C.C. Zou, W. Gong, D. Towsley, L. Gao, "The Monitoring and Early Detection of Internet Worms", *IEEE/ACM Trans. on Networking*, Vol. 13, Issue 5, pp. 961 - 974 Oct. 2005
- [14] T. Bu, A. Chen, S. Vander Wiel, T. Woo "Design and Evaluation of a Fast and Robust Worm Detection Algorithm" *IEEE INFOCOM 2006*
- [15] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, L. Peterson, "Characteristics of Internet Background Radiation", *ACM IMC*, October, 2004
- [16] A. Scherrer, N. Larrieu, P. Owezarski, P. Borgnat, P. Abry, "Non Gaussian and long memory statistical characterization of Internet traffic with anomalies", submitted to *IEEE Trans. on Dependable and Secure Computing*
- [17] P. Owezarski, "On the impact of DoS attacks on Internet traffic characteristics and QoS", *ICCCN 2005*, 17-19 October 2005
- [18] S. Floyd, V. Paxson, "Difficulties in simulating the Internet". *IEEE/ACM Trans. on Networking*, Vol. 9 , Issue 4, pp. 392-403, Aug. 2001
- [19] A. Dainotti, A. Pescapè, G. Ventre, "A Packet-level Characterization of Network Traffic", *CAMAD 2006*
- [20] A. Dainotti, A. Pescapè, P. Salvo Rossi, G. Iannello, F. Palmieri, G. Ventre, "An HMM Approach to Internet Traffic Modeling", accepted at *2006 IEEE GLOBECOM*
- [21] The CAIDA Dataset on the Witty Worm - March 19-24, 2004, Colleen Shannon and David Moore, <http://www.caida.org/passive/witty/>. Support for the Witty Worm Dataset and the UCSD Network Telescope are provided by Cisco Systems, Limelight Networks, the US Department of Homeland Security, the National Science Foundation, and CAIDA, DARPA, Digital Envoy, and CAIDA Members
- [22] WIDE Project: MAWI Working Group Traffic Archive <http://tracer.csl.sony.co.jp/mawi/>
- [23] <http://www.grid.unina.it/Traffic/>
- [24] D. Veitch, P. Abry, "A wavelet based joint estimator for the parameters of LRD", *IEEE Trans. Information Theory* Vol. 45, No.3, Apr. '99.
- [25] N. Weaver, I. Hamadeh, G. Kesidis and V. Paxson, "Preliminary Results Using Scale-Down to Explore Worm Dynamics", *ACM Workshop on Rapid Malcode (WORM)*, Oct. 2004.
- [26] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, C. Diot, "Packet-Level Traffic Measurements from the Sprint IP Backbone". *IEEE Network*, Vol. 17, Issue 6, pp. 6-16, Dec. 2003