Anomaly Characterization in Flow-Based Traffic Time Series

Anna Sperotto, Ramin Sadre, and Aiko Pras

University of Twente
Centre for Telematics and Information Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
P.O. Box 217, 7500 AE Enschede, The Netherlands
{a.sperotto, r.sadre, a.pras}@utwente.nl

Abstract. The increasing number of network attacks causes growing problems for network operators and users. Not only do these attacks pose direct security threats to our infrastructure, but they may also lead to service degradation, due to the massive traffic volume variations that are possible during such attacks. The recent spread of Gbps network technology made the problem of detecting these attacks harder, since existing packet-based monitoring and intrusion detection systems do not scale well to Gigabit speeds. Therefore the attention of the scientific community is shifting towards the possible use of aggregated traffic metrics. The goal of this paper is to investigate how malicious traffic can be characterized on the basis of such aggregated metrics, in particular by using flow, packet and byte frequency variations over time. The contribution of this paper is that it shows, based on a number of real case studies on high-speed networks, that all three metrics may be necessary for proper time series anomaly characterization.

1 Introduction

Attacks on our networks and server infrastructures are a growing source of concerns for network operators and users. They may be generated by both inexperienced script-kiddies and professional hackers, but in any case, attacks create unwanted traffic that can affect the performance and dependability of existing services. Therefore operators employ intrusion detection systems to identify and possibly filter suspicious traffic.

The constant increase in network traffic and the fast introduction of high speed (tens of Gbps) network equipment [16] make it hard to still employ traditional packet-based intrusion detection systems. Such systems rely on deep packet payload inspection, which does not scale well. In high speed environments, approaches that rely on aggregated traffic metrics, such as *flow-based* approaches, show a better scalability and therefore seem more promising. The advantage of *flow-based* approaches is that only a fraction of the total amount of data needs to be analysed. For the University of Twente, for example, we have estimated that the amount of flow data represents less than 1% of the amount of normal packet data.

A flow is defined as an unidirectional stream of packets that share common characteristics, such as source and destination addresses, ports and protocol type. In addition, a flow includes aggregated information about the number of packets and bytes belonging to the stream, as well as its duration. Flows are often used for network monitoring,

N. Akar, M. Pioro, and C. Skianis (Eds.): IPOM 2008, LNCS 5275, pp. 15–27, 2008.

permitting to obtain a real time overview of the network status; common tools for this purpose are Nfsen [5] and Flowscan [15], while the *de facto* standard technology in this field is Cisco Netflow, particularly its versions 5 and 9 [1,13]. The IETF IPFIX working group [14] is currently working on a standard for IP flow exporting, based on Netflow version 9.

Large networks, when creating flows, often apply *packet sampling* in order to make the approach even more scalable. In this case, only a percentage of the total number of packets passing through the monitoring point is considered in the flows. Statistical studies have been performed about correctness and precision of sampling strategies for Internet traffic [6] and high speed environments [7], as well as estimation of traffic flow characteristics from real sampled data [11]. These studies show that, despite the reduced amount of information, it is still possible to offer a correct statistical overview of the network status [6]. Packet sampling in flow creation is vastly deployed [12,17]. In particular, NetFlow relies on *systematic* sampling, where only 1 out of every n packets is considered for the accounting (1:n).

In the last years there has been an increasing interest in the application of flow-based techniques for anomaly and intrusion detection. The works of [8,9,10], which applies principal component analysis to traffic time series, and [19], which proposes a framework for network anomography, are examples of contributions in this field. Another example is provided by [2], which aims to detect worm spread in high speed network on a connection basis. In a similar environment, [3] addresses the problem of detecting DoS attacks and scans. In this case, the authors particularly focus on aggregated header information, as they can be exported by NetFlow (TCP flags). In addition, the presented approach is interesting because it explicitly addresses the problem of measure variation over time (with the use of value forecasting). In [4], the role of timely analysis of flow data is central. The author proposes a general purpose platform for parallel time-based analysis of flow information for attack detection, focusing in particular on DoS attacks (SYN-flood and web server overloading). From a network monitoring point of view, time series on flows, packets, and bytes are considered to be a useful tool: they permit to have a *dynamic* and *real time* overview of the network on the basis of the stream of information coming from the exporter [12,15].

In this paper, we investigate the use of traffic time series for identifying anomalies and detecting intrusions. In particular, we are interested in *whether it is necessary to consider 1) flows, 2) packets as well as 3) bytes time series, or whether it is sufficient to consider only one or two of these*. In addition, we want to know if the conclusion *also holds in the presence of sampling*. The novelty of our approach is that we rely on real case studies, performed in high-speed networks with links of 10 Gbps. Our measurements have been performed simultaneously on two different networks, the University of Twente (UT) and SURFnet, the Dutch research network, [18]. SURFnet applies 1:100 packet sampling during the flow creation.

The paper is organized as follow. Section 2 presents the measurement environment in which our analysis has been conducted. Sections 3 and 4 analyze anomalies in flow traces, focusing in particular on two real examples. Finally, Section 5 presents our conclusions.

2 Measurement Setup

The analysis presented in this paper has been conducted on flow traces collected at the University of Twente and on the SURFnet infrastructure [18]. In particular, the analyzed traces cover a period of time of two working days, namely between Wednesday August 1st 2007, 00:00 and Thursday August 2nd 2007, 23:59. The two networks have different sizes and coverage. The UT one is a $/16$ network providing connectivity to the employees and the students on the university buildings and the campus. SURFnet has national coverage and connects via optical path the most important research institutions in the Netherlands. Since SURFnet is also the UT network service provider, the majority of the incoming and outgoing UT traffic is routed through SURFnet. UT and SURFnet traces rely on a different measurement setup. Indeed, while UT processes all the packets passing through the measuring point, SURFnet applies a systematic sampling with ratio 1:100. In this paper, the real amount of traffic is estimated scaling all the measurements by a factor of 100.

Figure 1 shows the bytes traffic time series in the considered time frame. In this paper, all the time series have been created considering a time interval of 600 seconds, a good compromise between accuracy and number of samples. As expected, both networks show a clear night-day pattern, with peak of activity between 8:00 and 18:00 and with a minimum around 4:00. Around 16:00, on August 1st 2008, the amount of traffic on SURFnet drops abruptly. Since no error has been detected in our measuring setup, we suspect the down-peak to be caused by a flow creation and exporting failure in the SURFnet infrastructure, or, less likely, to a network hardware failure. Nevertheless, this event is not affecting our analysis. Table 1 presents the average, minimum and maximum traffic loads and the total data volume measured on the two networks during the observation period, together with the number of collected flows.

Table 1. Average, maximum and minimum traffic loads, data volume and number of flows during the period of observation on UT and SURFnet

	Avg Load	*Max Load*	*Min Load*	*Volume*	*Flows*
UT	652Mbps	1.01Gbps	259Mbps	21.65TB	982.7M
SURFnet	7.73Gbps	10.5Gbps	4Gbps	162.3TB	523.7M

During our monitoring time, UT seemed to be object of repeated and diverse attacks, even if apparently without real damage. Due to space constraints, we decide to concentrate our analysis on the following examples: `ssh` and `dns` traffic traces. The choice of these two specific sub-traces is due to the fact that, quite surprisingly, the `ssh` service resulted to be one of the major attack targets, both in intensity and in number of attacks. Similarly, by experience we noticed that `dns` tends to produce a quite regular traffic volume. This characteristic made quite easy to detect suspicious variation in traffic intensity. To properly evaluate if the observation in both networks are consistent, the `ssh` and `dns` traffic in SURFnet have been filtered in order to keep into account only the incoming-outgoing traffic from the UT network.

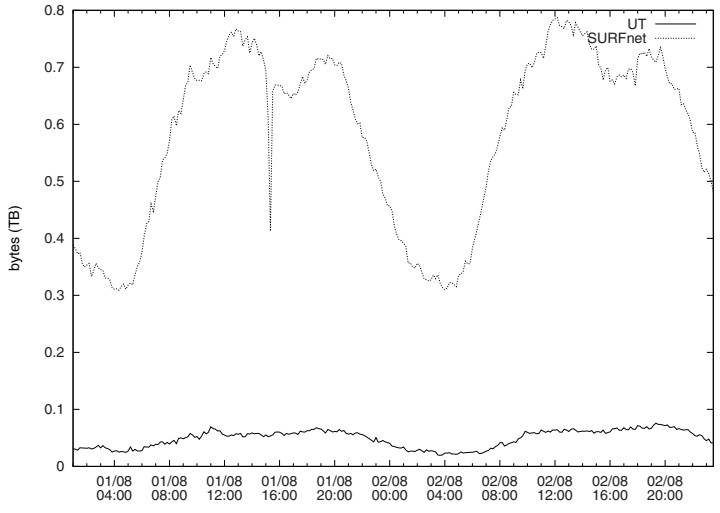

Fig. 1. Bytes time series, showing UT and SURFnet (estimated values) traffic

3 SSH Traffic

`Ssh` is one of the most common protocols to connect with remote machines. In general, it corresponds to the 1% of packets and the 1.2% of bytes of the total incoming-outgoing UT traffic.

3.1 Traffic Analysis

Figure 2 shows the byte traffic time series in the observation time frame. In the same graph, both UT and SURFnet (estimated) traffic volume are shown. It is possible to notice that the two measurements show the same trends, and only occasionally SURFnet strongly differs from UT. In general, the bytes trend in the observation period is quite irregular with sharp peaks and down-peaks. This situation is understandable because `ssh` can be used for both remote communications and file transfers. As a consequence, in the byte time series there is no clear evidence of attacks.

On the other side, looking at the packet time series (Figure 3), it is possible to notice that during the two days of observations, the UT network saw a massive increase of its `ssh` traffic. The time series is indeed characterized by sudden peaks during which the number of packets per time interval can raise of several millions. In some cases, we observe a difference of up to almost 8 millions packets. If we consider the flow time series, as in Figure 4, we can observe how the trend is also in this case characterized by peaks during which the number of flows per time interval raises form few thousand to half million. Again, the number of flows per time interval in SURFnet increases following the same behavior of the UT trace, despite the use of sampling. This phenomenon is particularly visible during the massive peaks, namely in the early morning of August 1st and in the late morning of August 2nd.

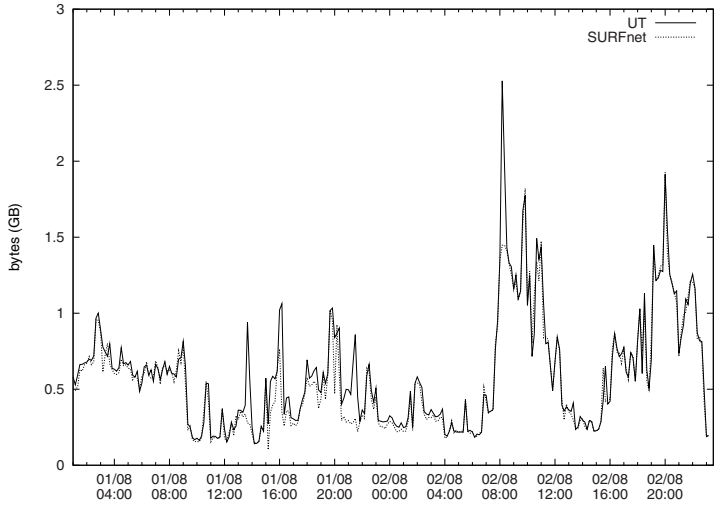

Fig. 2. Bytes time series, showing UT and SURFnet (estimated values) `ssh` traffic

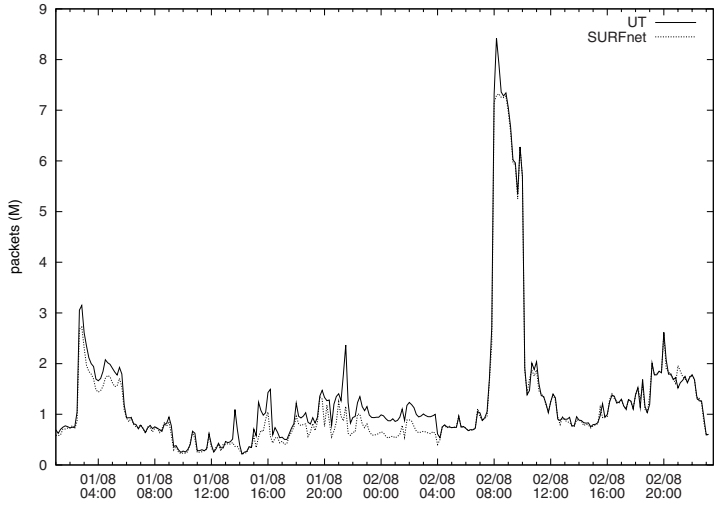

Fig. 3. Packets time series, showing UT and SURFnet (estimated values) `ssh` traffic

Summarizing, in the moments of major `ssh` activity, we observe a suspiciously high number of flows, matched by a very high number of packets, but with almost negligible amount of sent and received bytes. This suggest that the hosts involved are sending/receiving relatively small packets to many different hosts, scenario that suggests the possibility of a scan. A more detailed inspection of the trace shows indeed that few source hosts made the UT network object of massive `ssh` scans, during which the attackers were performing user-guessing on almost all the hosts in the UT network. It is

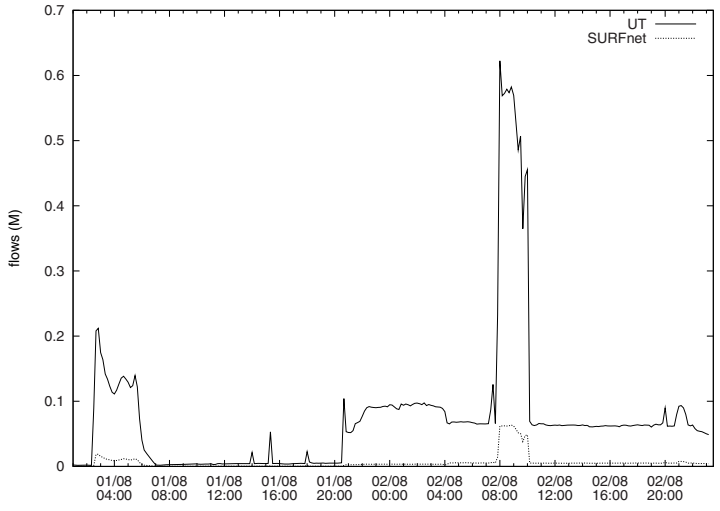

Fig. 4. Flows time series, showing UT and SURFnet `ssh` traffic

important to underline that it is the contemporaneous observation of all the three measure that permits to discriminate between normal and malicious traffic. For example, considering only peaks in the packet time series would not permit to distinguish file transfers from scanning activities. On the contrary, considering also bytes and flows would show that a file transfer has a different behavior from a scan, with peaks in the byte time series but not in the flow one. The traffic characteristics during the peaks made the `ssh` traffic trace worthy of deeper analysis. In the following, we concentrate only on the peak in the time frame from 7:50 to 10:10 on August 2nd, when the number of flows per time interval rises up to a maximum of almost 600000 flows (*(`ssh`) anomalous time frame*).

3.2 Normal vs Anomalous Traffic

The following analysis proves indeed that the previously identified peak is due to an attack. In order to characterize the network behavior during the anomaly, we need to compare it with a second observation time frame, that will provide us an overview of the network during a not suspicious interval. The second time windows span over a period of 2 hours, between 8:00 and 10:00 of August 1st. During this time frame, we are not observing any fast variation of the flow frequency. Since we are interested in scans and we are assuming that `ssh` scans produce variation in the flow frequency, we also assume the second time frame to be an example of *normal* network behavior (*normal time frame*).

Looking at the number of active hosts in the anomalous and normal time frames, Table 2 shows that the normal time frame is characterized by a balanced number of sources and destinations, both in UT and SURFnet. On the contrary, in the anomalous time frame, we can observe an increased number of destinations, several times bigger

Table 2. Number of distinct source and destination addresses during the anomalous and normal time frames in the UT and SURFnet traces

	Anomalous time frame		*Normal time frame*	
	Sources	*Destinations*	*Sources*	*Destinations*
UT	2763	65342	629	647
SURFnet	597	3020	192	192

than the number of sources. The number of destination hosts in the UT trace suggests that the scan covers the entire UT network (that is, as reported in Section 2, a $/16$ network), while the increased number of source hosts is an effect of the scanning activity (some of the destination hosts react to the probes). A similar trend is visible in SURFnet.

The study of the top active sources w.r.t. the number of originated flows shows that the anomalous time frame is dominated by the presence of three major senders, that caused the attack. Table 3 shows how the traffic, expressed in flows, packets and bytes, is distributed with respect to the sources during the anomalous time frame. Together, the three most active sources are responsible for the 98 - 99% of the total amount of flows in both UT and SURFnet. All the three hosts were scanning the UT network. As already suspected during the time series analysis, also the packet repartition is unbalanced towards the major senders (responsible of $\sim 70\%$ of the packets in both UT and SURFnet). Finally, it is important to notice that the scan *does not* deeply affect the bytes distribution: the 75% and the 69% of the bytes volumes respectively in UT and SURFnet is still due to normal traffic.

In order to give a visual representation of the network behavior during the anomalous and normal time frame, the scatter-plot in Figure 5 is presented. A time interval is characterized by a number of packets, bytes and flows. Let us suppose to assign to each measure an axis in a 3D space and plot each time interval as a point in this space. Figure 5 shows a representation of the anomalous and normal time frame. In the case of the anomalous time frame, also the projections on the planes are plotted. The graph permits to see that points belonging to the normal time frame tend to group together in a part of the space characterized by relatively small number of packets and bytes. Moreover, the time intervals in this group show a very low number of flows. On the contrary, the spatial disposition of the anomalous time frame describe a totally different behavior. Also in this case, the time intervals during the anomaly tend to be spatially close. This is an

Table 3. Percentage of flows, packets and bytes for the attackers and the not suspicious hosts during the `ssh` anomalous time frame

	Flow Percentage		*Packets Percentage*		*Bytes Percentage*	
	UT	*SURFnet*	*UT*	*SURFnet*	*UT*	*SURFnet*
SSH TOP 1	82.6%	89.5%	65.7%	71.2%	22.3%	28.1%
SSH TOP 2	13.5%	9.2%	6.7%	7.3%	2.3%	2.8%
SSH TOP 3	2.1%	0.3%	0.4%	0.3%	0.1%	0.1%
SSH OTHERS	1.8%	1%	27.2%	21.2%	75.3%	69.0%

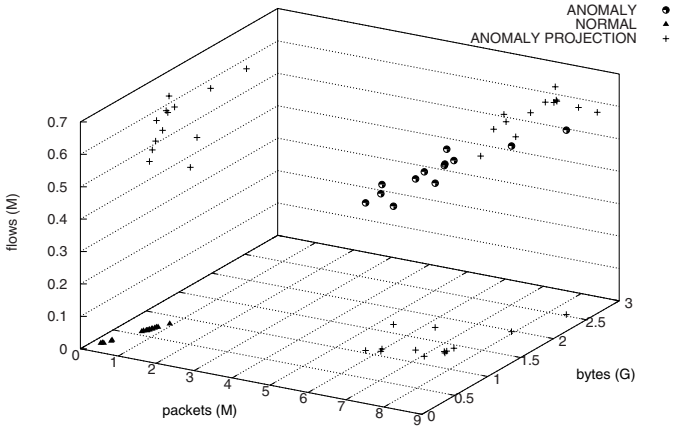

Fig. 5. `ssh` anomalous and normal time frame space disposition (UT trace)

indication of the fact that they share common features. In addition, as emphasized by the projections, points in this group present high values of the coordinates x (packets) and z (flows), while only few cases show a massive byte volume (y axis). Most importantly, the two groups are spatially distant to each other, confirming that anomalous and normal time intervals show clearly detectable differences.

4 DNS Traffic

`Dns` is the second trace we analyze in this paper. Commonly, `dns` is responsible of the less than 1% of the incoming-outgoing data volume at the UT network.

4.1 Traffic Analysis

In Section 3, `ssh` traffic seems to suggest that the flow frequency analysis can easily enlighten the presence of anomalies. Unfortunately, this hypothesis does not hold for `dns` traffic. As it is possible to see in Figure 6, the number of flows per time interval is almost constant during the entire observation period and nothing would suggest the presence of an anomaly.

The situation appears to be different if we are interested not in the flow time series but in the packet and byte ones. Figures 7 and 8, indeed, show that in the time windows between 1:40 and 7:00 am on August 1st, the UT network saw a massive increase in the volume of `dns` traffic, both in packets and in bytes. In particular, both measures raise abruptly from few thousands to millions (between 10 to 28 millions in a time interval). The SURFnet trace shows the same behavior, even in presence of sampling.

The just described anomaly is unnoticeable if only the flow time series is taken into account. This observation is particularly relevant because it witness how flow frequency variation is not expressive enough to characterize anomalies. By definition,

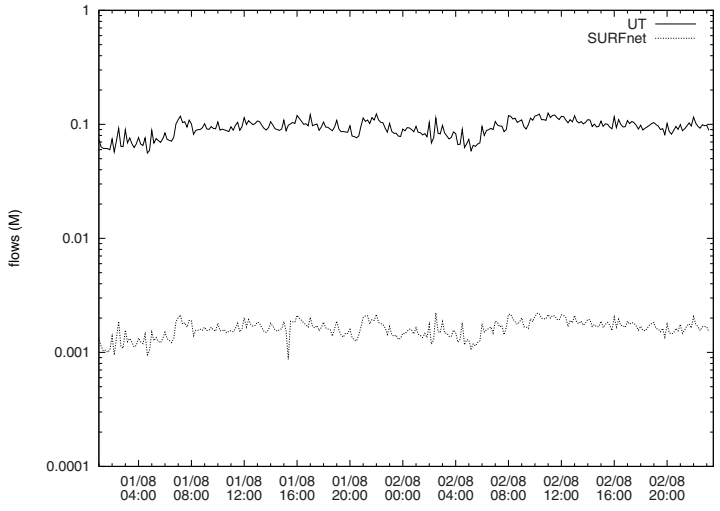

Fig. 6. Flows time series, showing UT and SURFnet `dns` traffic (in logarithmic scale)

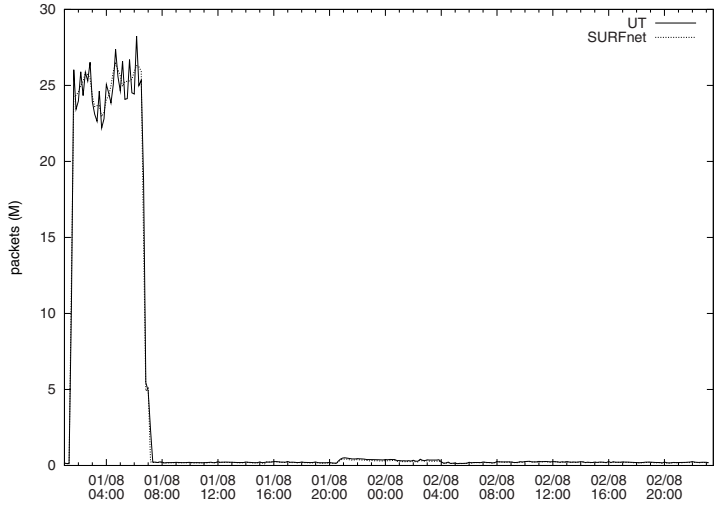

Fig. 7. Packets time series, showing UT and SURFnet (estimated values) `dns` traffic

`dns`traffic produces quite small UDP packets during the query process and it relies on TCP only in case of databases updates. Since the analysis of the protocol repartition during the anomaly shows that the 99.7% of the flows are UDP and they are responsible of the 99.9% of the bytes volume, we can exclude that the anomalies is caused by a database update. Under this consideration, we proceed for a more detailed analysis of the anomaly.

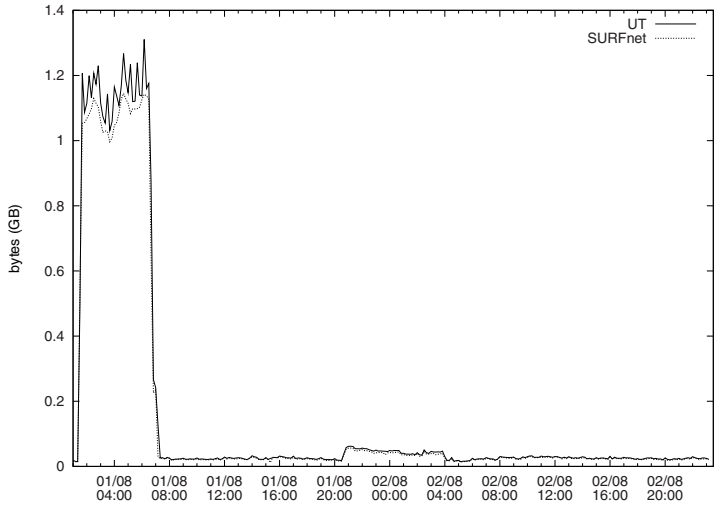

Fig. 8. Bytes time series, showing UT and SURFnet (estimated values) `dns` traffic

4.2 Normal vs Anomalous Traffic

As already for `ssh`, a not-anomalous interval has been chosen for sake of comparison. The *`dns` normal time frame* spans between 12:00 and 17:00 of August 1st. The large amount of bytes sent depicts a different scenario compared to the one presented in Section 3: the sharp variation in the byte and packets time series, together with the use of a large percentage of UDP packets suggests indeed the possibility of a DoS against a few number of destination hosts. The study of the anomalous time frame w.r.t the volume of byte sent clearly show the prevalence of three sources. Far away from the scenario of the `ssh` anomaly, the three sources are creating in average less that 300 flows each, being in this way responsible of only the 0.003% of the total UT flows. On the other side, each one of the major sources generates a packets volume almost 50 times bigger than all the other sources together. The proportion in the case of bytes is 20. SURFnet shows the same proportions. More generically, as it is possible to see in Table 4 the top senders host are responsible of more than 99% of the packets and the 98% of bytes in both UT and SURFnet traces. A deeper analysis of the traces shows that the three major sources share a single destination, towards which 33GB of data have been sent during the entire anomalous time frame (with packets of constantly exactly 46B in size). This configuration support the thesis that the destination host has been victim of a Distributed DoS targeting the `dns` service.

As previously in Section 3, a 3D representation of the anomalous and normal time frames is presented in Figure 9. Also in this case, the spatial disposition of the points in the two groups confirms the diversity between anomalous and normal time intervals. Points in the normal time frame show a relative variability in the number of flows, but almost no changes in the number of packets and bytes. On the contrary, the points in the anomalous group are characterized by large x and y coordinates (packets and bytes).

Table 4. Percentage of flows, packets and bytes for the attackers and the not suspicious hosts during the `dns` anomalous time frame

	Flow Percentage		*Packets Percentage*		*Bytes Percentage*	
	UT	*SURFnet*	*UT*	*SURFnet*	*UT*	*SURFnet*
DNS TOP 1	0.01%	0.14%	35.3%	35.3%	34.9%	34.8%
DNS TOP 2	0.01%	0.15%	32.6%	32.6%	32.3%	32.5%
DNS TOP 3	0.01%	0.14%	31.4%	31.4%	31%	31%
DNS OTHERS	99.97%	99.56%	0.7%	0.7%	1.8%	1.7%

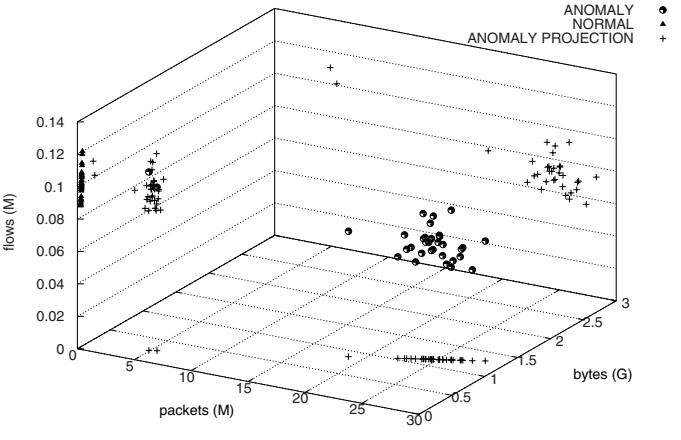

Fig. 9. `dns` anomalous and normal time frame space disposition (UT trace)

Only two time intervals during the anomalies are distant from the majority: they show indeed a relatively small number of packets and bytes. Nevertheless, the xy-projection of the anomaly confirms that this points are in any case anomalies. All the points in the anomalous time frame, with no exception of the two just described, belong to the same straight line. This is a consequence of the fact that the attackers were flooding the victim with fixed size packets. As final observation, in the graph it is possible to see that the number of flows during the anomalous and normal time frames does not differ enough to detect the ongoing attack, confirming the observation about the flow time series.

5 Conclusions

An important contribution of this paper is that our conclusions are based on extensive measurements on real, high speed networks, with line speeds of 10 Gbps. Our analysis confirm previous findings, that indicate that flows contain sufficient information to detect network intrusions. In particular, our study investigated whether flow, packet and byte time series are all needed to identify intrusions, or whether it is sufficient to consider only one or two of these metrics. Detailed analysis of two anomalies brought us to

the conclusion that, to correctly identify suspicious traffic in general, all three metrics should be taken into consideration.

Our analysis also showed that, for certain classes of attacks, the choice to monitor only a single metric may still be sufficient. This is for example the case in our flow time series for `ssh` traffic. On the other hand, such choice entails the risk of hiding other attacks. This is, for example, the case for the `dns` DoS attack, which does not appear in the flow time series. Therefore it is important to observe flow, packet as well as byte time series variation, to properly characterize anomalies.

Our study proves that this conclusion also holds in the presence of sampling. Sections 3 and 4 showed that the sampled traces closely approximate the non-sampled traces, which means that accurate anomaly detection is possible even in case of sampling. This observation suggest that the development of scalable, but still accurate intrusion detection solutions is possible.

Finally, Sections 3.2 and 4.2 ouline directions for future work. Normal traffic and traffic generated during an anomaly show a clear spatial division. This ensures us that modelling network behaviours is possible. Our future studies will deal with the creation of models suitable, first of all, for the problem of detection and, at the same time, effective for real time analysis of high speed networks.

Acknowledgments. This research has been supported by the EC IST-EMANICS Network of Excellence (#26854). We also would like to thank Daan van der Sanden for his valuable help in the traces analysis.

References

1. Claise, B.: Cisco Systems NetFlow Services Export Version 9. Request for Comments: 3954, IETF (October 2004)
2. Dubendorfer, T., Plattner, B.: Host behaviour based early detection of worm outbreaks in internet backbones. In: WETICE 2005: Proc. of the 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise, pp. 166–171. IEEE Computer Society, Washington (2005)
3. Gao, Y., Li, Z., Chen, Y.: A dos resilient flow-level intrusion detection approach for high-speed networks. In: ICDCS 2006: 26th IEEE International Conference on Distributed Computing Systems, pp. 39–39 (2006)
4. Munz, G., Carle, G.: Real-time analysis of flow data for network attack detection. In: IM 2007: 10th IFIP/IEEE International Symposium on Integrated Network Management, 2007, pp. 100–108 (2007)
5. Haag, P.: Nfsen: Netflow sensor (April 2008), `nfsen.sourceforge.net`
6. He, G., Hou, J.C.: An in-depth, analytical study of sampling techniques for self-similar internet traffic. In: ICDCS 2005: Proc. of the 25th IEEE International Conference on Distributed Computing Systems, pp. 404–413. IEEE Computer Society, Los Alamitos (2005)
7. Izkue, E., Magaña, E.: Sampling time-dependent parameters in high-speed network monitoring. In: PM2HW2N 2006: Proc. of the ACM international workshop on Performance monitoring, measurement, and evaluation of heterogeneous wireless and wired networks, pp. 13–17. ACM, New York (2006)
8. Lakhina, A., Crovella, M., Diot, C.: Characterization of network-wide anomalies in traffic flows. In: IMC 2004: Proc. of the 4th ACM SIGCOMM conference on Internet measurement, pp. 201–206. ACM, New York (2004)

9. Lakhina, A., Crovella, M., Diot, C.: Diagnosing network-wide traffic anomalies. In: SIGCOMM 2004: Proc. of the Conference on Applications, technologies, architectures, and protocols for computer comm., pp. 219–230. ACM, New York (2004)
10. Lakhina, A., Papagiannaki, K., Crovella, M., Diot, C., Kolaczyk, E.D., Taft, N.: Structural analysis of network traffic flows. SIGMETRICS Perform. Eval. Rev. 32(1), 61–72 (2004)
11. Yang, L., Michailidis, G.: Sampled based estimation of network traffic flow characteristics. In: INFOCOM 2007. 26th IEEE International Conference on Computer Communications, pp. 1775–1783. IEEE, Los Alamitos (2007)
12. Cisco IOS NetFlow (April 2008), http://www.cisco.com/go/netflow
13. Cisco IOS NetFlow Configuration Guide (April 2008), http://www.cisco.com
14. IP Flow Information Export Working Group (April 2008), http://www.ietf.org/html.charters/ipfix-charter.html
15. Plonka, D.: Flowscan (April 2008), http://www.caida.org/tools/utilities/flowscan/
16. Internet2 NetFlow: Weekly Reports. netflow.internet2.edu/weekly (April 2008)
17. sFlow (April 2008), http://www.sflow.org
18. SURFnet (April 2008), http://www.surfnet.nl
19. Zhang, Y., Ge, Z., Greenberg, A., Roughan, M.: Network anomography. In: Proceedings of the Internet Measurement Conference 2005 on Internet Measurement Conference, pp. 317–330. USENIX Association (2005)