Difficulties in Modeling SCADA Traffic: A Comparative Analysis

Rafael R. R. Barbosa, Ramin Sadre, and Aiko Pras

University of Twente
Design and Analysis of Communication Systems (DACS)
Enschede, The Netherlands
{r.barbosa,r.sadre,a.pras}@utwente.nl

Abstract. Modern critical infrastructures, such as water distribution and power generation, are large facilities that are distributed over large geographical areas. Supervisory Control and Data Acquisition (SCADA) networks are deployed to guarantee the correct operation and safety of these infrastructures. In this paper, we describe key characteristics of SCADA traffic, verifying if models developed for traffic in traditional IT networks are applicable. Our results show that SCADA traffic largely differs from traditional IT traffic, more noticeably not presenting diurnal patters or self-similar correlations in the time series.

1 Introduction

Modern critical infrastructures, such as water distribution and power generation, are large facilities that are distributed over large geographical areas. Supervisory Control And Data Acquisition (SCADA) networks are deployed to continuously monitor these infrastructures in order to guarantee correct operation and safety. Originally, SCADA networks were isolated networks running proprietary protocols, but there is an increasing trend toward the usage of IP protocols and the interconnection with other networks and even the Internet.

Intuitively, we expect SCADA to present traffic patterns much different to those of "traditional" Information Technology (IT) networks. This is due to a number of reasons. First, SCADA networks are expected to be more stable over time, in the sense that new nodes are not expected to join or leave frequently. Second, traditional networks usually support a multitude of protocols, such as HTTP, instant messaging and Voice over IP, while the number of services in SCADA networks is expected to be more limited. Finally, most of the SCADA traffic is expected to be generated in a periodical fashion, due to the polling mechanism used to gather data. In consequence, traffic patterns should not be so dependent on human activity as in traditional IT networks.

Apart from the assumptions given above, not much more is publicly known about the behavior of SCADA traffic. This is partly caused by the sensitivity of the data. In fact, publications on SCADA networks generally do not rely on empirical data as obtained from real-world measurement [1,2,3]. In contrast,

traditional networks have been intensively studied, sometimes leading to surprising insights. As an example, we refer to the seminal work in [4] and [5] on the self-similar nature of network traffic and, connected to that, to studies on the presence of long-range dependency and heavy-tailed distributions [6,7,8,9]. The research has resulted in models and tools employed in, for example, the design and dimensioning of network equipment and the parametrization of management algorithms. Naturally, the question arises whether the existing models are also valid for SCADA networks.

The goal of this paper is *to verify if models used to describe traditional network traffic can also be applied to SCADA traffic*. We achieve this by comparing a traditional IP traffic trace with *real-world* SCADA measurements done by us. However, network behavior can be compared in a virtually infinite number of ways, starting from the above mentioned characteristic of self-similarity to topological properties [10] and application specific aspects [11]. In order to provide information that is of interest for a wide range of readers, we base our analysis in this first work on a list of "invariants", i.e., behaviours that are empirically shown to hold for a wide range of environments, proposed in the well known paper of Floyd and Paxson [12]. We revise this list and test our datasets for the invariants applicable to our context.

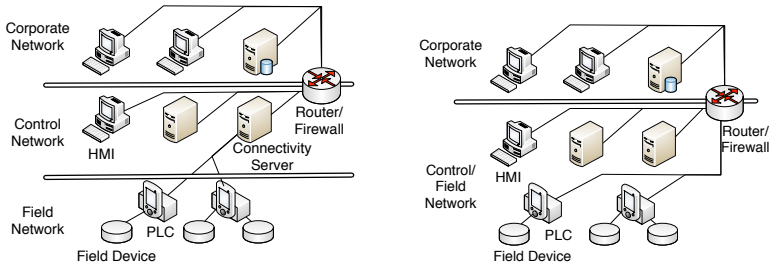
In a separate, but closely related work, we perform a series of tests to characterize SCADA traffic at the IP level, while drawing a comparison with Simple Network Management Protocol (SNMP) traffic [13]. Our analysis confirms that most hosts (including user workstations) generate data in a periodical way, resulting in a remarkably constant traffic time series. Surprisingly, we observe that changes in the IP level connectivity matrix are common.

The rest of this paper is organized as follows. In Section 2, we describe the datasets used in this paper. In Section 3, we give a short description of the invariants and we briefly explain how the tests are performed. The results are presented in Section 4. Finally, conclusions are given in Section 5.

2 Datasets

The datasets that we use in this paper consist of four network packet traces in pcap format [14], collected at three different locations: two water treatment and distribution facilities that use SCADA networks and one research institute network with "ordinary" IP traffic. From the pcap traces we generate flow information by aggregating packets that are no more than 300s apart, based on the traditional 5-tuple of protocol number, source and destination IP addresses and port numbers. In this section, we give more insight into the data.

The two SCADA locations have different topologies, as shown in Figure 1. Both topologies have a *corporate network* that does not have direct access to the other parts of the network and is, in general, connected to the Internet. In the three-layer topology (Figure 1a) , the remaining part of the network consists of the *field network* and the *control network*. The field network contains the Programmable Logic Controllers (PLC) and the Remote Terminal Units (RTU)



(a) Three-layer topology (b) Two-layer topology

Fig. 1: SCADA topologies of the monitored networks

Name	Number of hosts	Duration	Average pkts/s	Average KBytes/s
2layer	45	13 days	504.1	82.5
3layer-control	14	10 days	28.7	5.1
3layer-field	31	10 days	75.7	28.2
IT	100	7.5 days	81.9	65.3

Table 1: Datasets overview

that monitor (and possibly issue commands to) the field devices. The control network contains several servers with different purposes, such as automatically polling of field nodes and performing the access control; and the Human-Machine Interfaces (HMI). The latter are operator workstations that provide an user interface to the field nodes. The communication between the control network and the field network passes through a single node, the connectivity server. In contrast, there is no such explicit (physical) separation between the control network and the field network in the two-layer topology (Figure 1b).

For the SCADA location following the 2-layer topology, we have captured the traffic in the joint control/field network. We refer to the collected dataset as *2layer* in the following. For the 3-layer SCADA location, we have captured the traffic in the control network as well as in the field network. The so obtained two packet traces are referred to as *3layer-control* and *3layer-field*. In both locations, the data capture was done through a switch's *mirror port*, that replicated all traffic in a given network. No data loss was reported. Finally, we have ignored the traffic in the corporate networks since they do not transport SCADA traffic.

In order to provide a comparison with a traditional IT environment, we have selected a publicly available traffic trace from the network of an educational organization: Location 6 from [15]. The organization is relatively small with around 36 employees and 100 students. Its network is comparable to the above SCADA networks in the number of hosts as well as in the average bandwidth and, hence, is an adequate candidate for the following studies. We use only a portion of the available data, approximately the first 7.5 days of the trace. We refer to this dataset as *IT*. An overview of all four datasets is given in Table 1.

3 Invariants

In [12], seven invariants in Internet traffic are presented. Not all of them are suitable for the datasets considered in this paper. In Sections 3.1 through 3.3, we give a short description of those four invariants that we test in Section 4 and we briefly explain how the tests are performed. In Section 3.4 we discuss the remaining three invariants and the reasons why we have not considered them.

3.1 Diurnal patterns of activity

Network activity is strongly correlated with human activity. As a consequence, it starts increasing around 8–9 AM local time, peaks around 11 AM and 3–4 PM and decreases as business day ends at 5 PM. Moreover, the amount of traffic during the weekends tends to be considerably smaller than during week days. In order to verify if SCADA traffic also follows this pattern, we plot time series for three different measures: the number of active flows, *packets/sec* and *bytes/sec*.

3.2 Self-similarity

Self-similarity is the quality that the whole resembles its parts. In network traffic, it can be observed as bursty periods being present at different timescales, from milliseconds to a few hours. This property violates the assumptions of traditional Markovian modeling that predicts that longer-term correlations are weak. Since the initial findings in the early 90's [4,5], self-similarity of network traffic has remained an active field of research (see, e.g., [9]).

For this paper, we have decided to employ three popular visual methods to test self-similarity [4,6]: the R/S analysis, variance-time plots and periodograms. The visual representation of their results allows to detect anomalies and to estimate the degree of self-similarity in the data:

R/S analysis: For a given set of observations $X = X(t), 0 < t \leq N$, consider a subset with starting point t_i and size n. Let $\overline{X}(t_i, n)$ and $S(t_i, n)$ be, respectively, the mean and the standard deviation of a subsample of X calculated over the interval $[t_i, t_i+(n-1)]$. The *rescaled adjusted range* plot (or *R/S pox diagram*) can be obtained by dividing a set of observations X into K non-overlapping subsets of size N/K with starting points $t_i = i(N/K) + 1$. One selects logarithmically spaced values of n and plots $\log(R/S(t_i, n))$ as a function of $\log(n)$, where R/S is the *R/S statistic*. The Hurst parameter can be estimated from the slope of a line fitted to the resulting curve.

Variance-time plots: Self-similar time series do not become "smoother" at larger time scales, i.e., the variance decreases slowly for increasing aggregation levels. Let $X^{(m)}$ be the aggregated process, defined as $X^{(m)}(t) = m^{-1} \sum_{t=1}^{t+(m-1)} X(t)$. The variance-time plot shows the variance of the aggregated process, $S^2(X^{(m)})$ versus the aggregation level m in a log-log scale. A line is least-squares fitted to the resulting curve, ignoring small values of m. A slope $-1 \leq \beta \leq 0$ suggests self-similarity, and the Hurst parameter can be estimated as $H = 1 - \beta/2$.

Periodograms: The last method consists of fitting a least-squares line to the low-frequency part of a periodogram, typically the lowest 10%. The Hurst parameter can be estimated as $H = (1 - \beta)/2$, with β being the slope of the fitted line.

3.3 Log-normal connection sizes and Heavy-tail distributions

Log-normal distributions are a good fit to the body of connection size distributions, while the tails of network-activity related distributions are often heavy-tailed. Since the original list of invariants was published, a debate started over which of these models better describe connection size distributions: heavy-tail (e.g., [6]) or log-normal (e.g., [7]). Recently, Gong et al. [8] argued that there is never sufficient data to support any analytical form summarizing the tail behavior, therefore the research efforts should focus instead on studying the complex nature of traffic generation and its implications.

In this work, we do not attempt to fit our measurements to theoretical distributions. We simply show, through widely used Complementary Cumulative Distribution Functions (CCDFs) [7], that measurements from the *IT* dataset generally match the results reported in the literature and point out the differences to the connection size distributions in SCADA networks.

3.4 Invariants not tested in this work

In addition to the above four invariants, [12] also defines three invariants that we do *not* further study in this paper for reasons explained in the following:

Session arrivals: A "session" refers to the period of time a human uses the network for a specific task. There is evidence that session arrivals are well-modeled by a Poisson process, e.g., FTP, TELNET [5] and HTTP [16]. Since the concept is highly protocol specific, it is hard to develop a general method to group network packets to sessions. This is especially true for our SCADA datasets, as most of the protocols are closed. Hence, we do not attempt to test this invariant in this work. Note that *flows* are *not* well-modeled by a Poisson process.

Telnet packet generation: Packets generated by keystrokes, e.g., in a Telnet session, obey a Pareto distribution. Since this invariant mostly concerns human behavior and a single specific protocol, we have not considered it in this work.

Characteristics of the global topology: Some behaviors appear due to characteristics of the Earth. For example, the delay in inter-continental connections is bounded by the propagation delay. Such characteristics are not relevant for the relatively small networks considered in this paper.

4 Analysis results

In this section we discuss the results of our analysis regarding the four selected invariants.

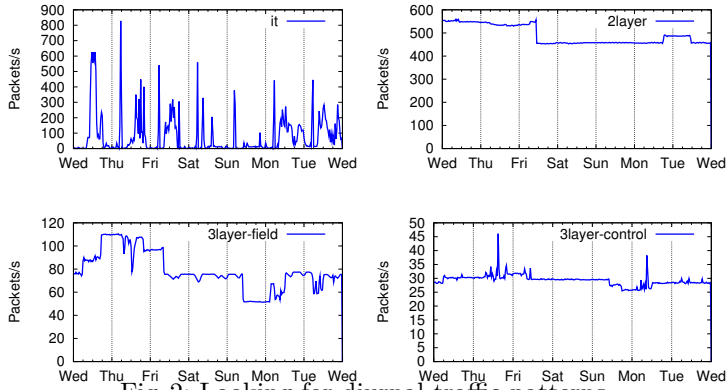


Fig. 2: Looking for diurnal traffic patterns

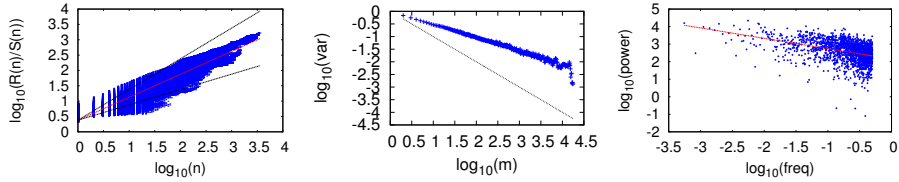
4.1 Diurnal Patterns of Activity

Diurnal patterns in network activity are widely reported in the literature [12]. In contrast, most of the traffic of a SCADA environment is generated periodically by the polling mechanism used to retrieve data, and as a consequence, it should have a very regular throughput. To verify this, we plot three different time series: *packets/s*, *bytes/s* and *number of active flows*, calculated over 30-minute bins for our four datasets. To ease the comparison, we align the time series based on weekdays. Figure 2 show the results for *packets/s*. The results for the other metrics are analogous, thus not shown due to space constraints.

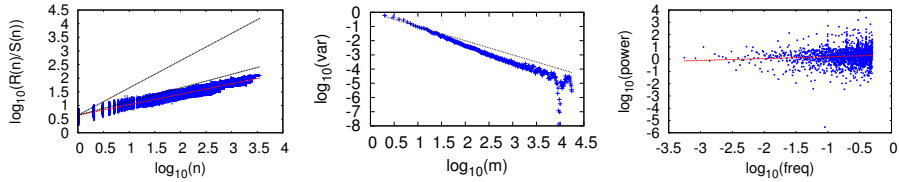
As can be seen, the SCADA traffic does not present day and night patterns. Instead, all time series remain stable over large periods of time, to which we refer as *baselines*. Note, however, that the throughput is not constant. Notably, datasets *2layer* and *3layer-field* present a considerable drop in the packet rate at around Friday noon and Sunday noon respectively. Such stability combined with the fact that most sources generate traffic in a periodical way [13] indicates that *ON/OFF* models might provide a good approximation for the general shape of the time series.

A closer inspection of the data reveals three major causes for the deviations from the baseline: (i) the start or end of flows with large throughput, (ii) the increase (or decease) in the rate in which variables are pooled and (iii) the increase (or decrease) in the number of variables pooled. We speculate that the changes are mostly caused by certain changes in the physical process that the SCADA systems control, e.g., tanks becoming full or an increase in the water demand. Another possible cause is a manual access to the PLCs, for either retrieving data or uploading a new configuration. Further research is necessary to establish if these changes can be predicted.

As expected, the *IT* dataset shows diurnal patterns of activity, with lower throughput during the nights and weekends. The daily peaks seen in the early morning (around 5.25 AM) are caused by a single large flow between the same



(a) R/S diagram (IT trace) (b) Variance-time (IT trace) (c) Periodogram (IT trace)



(d) R/S diagram (3c trace) (e) Variance-time (3c trace) (f) Periodogram (3c trace)

Fig. 3: Self-similarity tests on the IT trace and the 3layer-control (3c) trace

two hosts. We assume it to be related to some automated activity, such as backup, but we did not attempt to verify which.

4.2 Self-Similarity

One of the requirements for a random variable to be self-similar is that it must be wide-sense stationary [4], which implies, among other things, a constant mean over time. Therefore, due to the diurnal patterns of activity, network traffic is not truly self-similar [8]. However, network measurements with durations up to a few hours *do* present self-similar time series [4,6]. Other sources of non-stationarity are singular events that cause drastic changes in the network behavior, such as a maintenance operation or changes in physical processes (see Section 3.2). For the following analysis, we have taken periods of a few hours from our datasets where the stationarity requirement is satisfied.

The self-similarity analysis is performed for the $pkts/s$ and $bytes/s$ time series with 100 millisecond bins for all datasets. The results for the $bytes/s$ are analogous to the ones from $pkts/s$ and, therefore, are omitted. Figure 3 depicts the R/S pox diagram, the variance time plots and the periodograms for *IT* in the first row (Figures 3a, 3b and 3c respectively) and *3layer-control* in the second row (Figures 3d, 3e and 3f respectively). The results for the other SCADA datasets are analogous, thus also omitted.

The R/S pox diagram of a self-similar random variable should have an asymptotic slope between 0.5 and 1 (represented by the black dotted lines). The slope is typically estimated by least-square fitting (represented by the red dotted line). It is clear from Figures 3a and 3d that *IT* presents self-similar behavior, while *3layer-control* does not. A comparable result is obtained using the variance-time

dataset	bytes/s			pkts/s		
	R/S	var-time	period	R/S	var-time	period
IT	0.73	0.72	0.79	0.75	[0.71-0.72]	0.79
2layer	0.17	[0.09,0.11]	0.13	0.17	[0.32,0.42]	0.22
3layer-control	0.38	[0.38,0.44]	0.43	0.39	[0.36,0.37]	0.44
3layer-field	0.02	[0.27,0.31]	0.29	0.44	[0.35,0.42]	0.04

Table 2: Hurst parameter estimations

plot test, where the slope of the resulting curve should be shallower than -1 (black dotted line). This test shows that the variance of the SCADA time series decays much faster than the expected for a self-similar process. In contrast, the *IT* dataset result is consistent with the traditional network measurements. The same conclusion can be drawn from the periodogram test. When applying this method, we obtain a estimative of $H = 0.79$ for the *IT* dataset and of $H = 0.44$ for *3layer-control*. Note that the Hurst parameter of a self-similar process should be in the interval $H \in [0.5, 1)$.

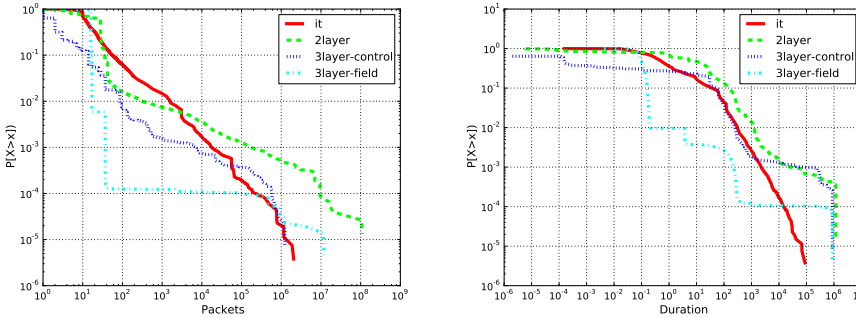
Table 2 summarize the results of our analysis, reporting the estimates for the Hurst parameter from the R/S analysis (*R/S*), variance-time plots (*var-time*) and periodograms (*period*). All estimates for the SCADA datasets indicate a non-self-similar behavior, although the estimates are not consistent between tests. In contrast, the *IT* dataset shows more consistent estimate of the Hurst parameter, which is in agreement with a self-similar behavior. Note also that, while the R/S analysis and periodograms yield a single estimate, the variance-time plots produce a small range of estimates. This happens because for both small and large aggregation levels m there is a considerable amount of variance that should not be taken into account when performing the least-square fit. In our analysis we remove up to 15% of either end of the variance-time plot to obtain the Hurst estimates.

4.3 Distributional Aspects of Connection Sizes

As explained in Section 3.3, there is a debate in the research community around which distribution best fits the tail behavior of connections sizes[1], heavy-tail (usually Pareto-distributed) or log-normal. We can illustrate both behaviors for the *IT* trace. In the case of the number of packets per flow, plotted in Figure 4a, the CCDF presents an almost constant slope, indicating that a Pareto model might provide a good fit. In the case of flow duration, plotted in Figure 4b, the behavior is closer to that of an log-normal distribution, with an increasing slope when approaching extreme values in the tail.

For the SCADA datasets, the results are not always conclusive. For instance, consider again the connection size in packets plotted in Figure 4a. The tail for dataset *2layer* could be modeled as Pareto, if one considers the tail to consist of values above 10^2. In the case of *3layer-control*, the CCDF presents large variations and cannot be approximated by either model. When considering duration,

[1] In this paper we use the terms connection and flow interchangeably. For the definition of flow we refer to Section 2.



(a) Packets CCDF (b) Flow duration CCDF

Fig. 4: Flow size Complementary CDF's

2layer and *3layer-control* CCDFs present different slopes at different ranges. See for instance the CCDF of *3layer-control* in Figure 4b. The slope is relatively small up to $10s$, it sharply increases in the interval $[10, 10^3]$ after which it sharply reduces. Finally, the tail of dataset *3layer-field* for both metrics is dominated by a small range of values, which produces the nearly vertical lines in both plots.

Irrespective of which is the best model to represent the connection size distribution, all datasets share a common characteristic: the connection size distribution is always *positively skewed*, i.e, it has a body containing the majority of the values in the distribution and a tail with extreme values in the right.

5 Conclusions

The goal of this paper was to verify if models used to describe traditional network traffic can also be applied to SCADA traffic. To this end, we have analyzed SCADA traffic traces collected at two water treatment and distribution facilities and compared their characteristics with those of traditional network traffic. Our analysis has been based on a list of network traffic invariants widely observed in network measurements.

We draw the following conclusions from our results. First, SCADA networks do not present the diurnal patterns of activity common to traditional IT networks, as most of the traffic is generated by automated processes with little human interaction. More important, self-similar correlations in the time series are not present. Our results suggest that simple *ON/OFF* models might provide a good approximation for the time series. Finally, neither heavy-tail nor log-normal models seem to provide a good fit for the connection sizes. In summary, our results indicate that the existing traffic models can not be easily applied to SCADA traffic.

To our best knowledge, we provide the first study on real-world SCADA traces in this paper. Since existing publications on SCADA networks generally do not rely on empirical data, we believe that our findings are a first step towards

constructing realistic SCADA traffic models to support future research in the area. In future work, we intend to extend our analysis of SCADA traffic, including the characterization of the flow arrival process and the extraction of periodical patterns.

References

1. Kobayashi, T., Batista, A., Brito, A., Pires, P.: Using a packet manipulation tool for security analysis of industrial network protocols. In: IEEE Conference on Emerging Technologies and Factory Automation (ETFA). (2007) 744–747
2. Cheung, S., Skinner, K., Dutertre, B., Fong, M., Lindqvist, U., Valdes, A.: Using model-based intrusion detection for SCADA networks. In: Proceedings of the SCADA Security Scientific Symposium, Citeseer (2007) 1–12
3. Valdes, A., Cheung, S.: Communication pattern anomaly detection in process control systems. In: Technologies for Homeland Security, 2009. HST '09. IEEE Conference on, IEEE, IEEE (May 2009) 22–29
4. Leland, W.E., Willinger, W., Taqqu, M.S., Wilson, D.V.: On the self-similar nature of Ethernet traffic. ACM SIGCOMM Computer Communication Review **25**(1) (January 1995) 202–213
5. Paxson, V., Floyd, S.: Wide area traffic: the failure of Poisson modeling. IEEE/ACM Transactions on Networking **3**(3) (June 1995) 226–244
6. Crovella, M., Bestavros, A.: Self-similarity in World Wide Web traffic: evidence and possible causes. Networking, IEEE/ACM Transactions on **5**(6) (1997) 835–846
7. Downey, A.: Lognormal and Pareto distributions in the Internet. Computer Communications **28**(7) (May 2005) 790–801
8. Gong, W.B., Liu, Y., Misra, V., Towsley, D.: Self-similarity and long range dependence on the internet: a second look at the evidence, origins and implications. Computer Networks **48**(3) (2005) 377 – 399 Long Range Dependent Traffic.
9. Loiseau, P., Goncalves, P., Dewaele, G., Borgnat, P., Abry, P., Primet, P.V.B.: Investigating Self-Similarity and Heavy-Tailed Distributions on a Large-Scale Experimental Facility. IEEE/ACM Transactions on Networking **18**(4) (August 2010) 1261–1274
10. Vázquez, A., Pastor-Satorras, R., Vespignani, A.: Large-scale topological and dynamical properties of the internet. Physical Review E **65** (2002)
11. Sadre, R., Haverkort, B.: Changes in the Web from 2000 to 2007. In: Managing Large-Scale Service Deployment. Proceedings of the 19th IFIP/IEEE International Workshop on Distributed Systems (DSOM 2008). Volume 5273 of LNCS., Springer (2008) 136–148
12. Floyd, S., Paxson, V.: Difficulties in simulating the Internet. IEEE/ACM Transactions on Networking **9**(4) (2001) 392–403
13. Barbosa, R.R.R., Sadre, R., , Pras, A.: A First Look into SCADA Network Traffic. In: Network Operations and Management Symposium (NOMS). (2012) To appear.
14. Jacobson, V., Leres, C., McCanne, S., et al.: Tcpdump (1989)
15. Barbosa, R.R.R., Sadre, R., Pras, A., van de Meent, R.: Simpleweb/university of twente traffic traces data repository. Technical report, Centre for Telematics and Information Technology, University of Twente (April 2010)
16. Nuzman, C., Saniee, I., Sweldens, W., Weiss, A.: A compound model for TCP connection arrivals for LAN and WAN applications. Computer Networks **40**(3) (October 2002) 319–337