A First Look into SCADA Network Traffic

Rafael Ramos Regis Barbosa, Ramin Sadre, and Aiko Pras
Design and Analysis of Communications Systems (DACS)
University of Twente, The Netherlands
Email: {r.barbosa, r.sadre, a.pras}@utwente.nl

***Abstract*—Supervisory Control and Data Acquisition (SCADA) networks are commonly deployed to aid the operation of critical infrastructures, such as water distribution facilities. These networks provide automated processes that ensure the correct functioning of these infrastructures, in a operation much similar to those of management operations found in traditional Internet Protocol (IP), in particular the Simple Network Management Protocol (SNMP). In this paper we provide a first look into characteristics of SCADA traffic, with the goal of building an empirical foundation for future research, and investigate to what extent the SCADA traffic patterns are similar to SNMP.**

I. Introduction

Critical infrastructures, such as power grids and water distribution facilities, comprise a high number of devices over a large geographical area. Supervisory Control and Data Acquisition (SCADA) networks are implemented to coordinate and manage the actions of these devices. Workstations provide human operators with real-time information about the field process and with some control capabilities. Automated processes ensure that the infrastructure operates correctly and safely, by continuously polling field information and, eventually, sending control commands. Alerts are generated in case of severe problems, so human operators can intervene.

In fact, such operation is very similar to management operations found in traditional Internet Protocol (IP) networks, such as the ones provided by the Simple Network Management Protocol (SNMP) [1]. Probably the most widespread network management protocol, SNMP is used by automated applications to provide network managers with real-time information about the network. These applications ensure that the network infrastructure operates correctly, by continuously polling information from the devices. Alerts are also sent in case of problems, so that network managers can intervene.

Based on these similarities, we propose the following research question: *Is SCADA traffic similar to SNMP?* In [2], we show that some existing models created to describe *traditional* Information Technology (IT) traffic (e.g., self-similarity) do not apply to SCADA. The answer to this question would indicate if it is possible to apply part of the knowledge acquired in many years of SNMP research, in areas such as security and performance, to SCADA research. To address this problem we provide a first look into real-world SCADA traffic and draw a comparison with SNMP. We perform a series of tests with the objective to characterize traffic at the IP level.

978-1-4673-0269-2/12/$31.00

II. Related Work

Research in SCADA seems to be mostly dedicated to security issues. In [3], the authors identify general threats regarding SCADA environments and provide a list of research challenges in the area. The lack of Intrusion Detection Systems (IDS) for SCADA is one of the challenges addressed [4], [5]. In [4], an IDS is proposed that is mainly based on the Modbus protocol specification, but also assumes regularity and stability in regard of topology, communication and configuration. A different approach is used in [5], where the communication patterns are the basis of the proposed IDS. The problem of these approaches is that they assume certain patterns without having empirical results to support them. We argue that measurements should be an essential part of IDS SCADA research, as they allow the validation of the traffic models used.

SNMP have for long been the subject of numerous books (e.g., [6]), papers (e.g., [7]) and software development (e.g., [8]). However, it has been observed in a recent survey by Andrey et al. [7] that the research community does not agree in well-defined and commonly agreed criteria to evaluate SNMP. A first effort in that direction is made in [9], where SNMP data is collected and analyzed with the goal of revealing traffic patterns in real networks. This analysis is extended in [10], with the addition of several datasets.

III. Datasets

In this study we analyze nine different datasets: six SNMP and three SCADA *tcpdump/libpcap* [11] traces. The SNMP traces are the same studied in [9], and its extension [10]. As such, we adopt the same naming scheme: the first number in the name identifies the location where the trace was collected and the second number gives the trace number, e.g., the dataset *l01t02* is the second trace from location 1. The single character at the beginning of the name either indicates a SCADA trace (letter "s") or a SNMP trace (letter "l").

Some of the datasets contain large *gaps*, i.e. periods without any traffic. For those traces, we restrict our analysis to the longest period without any gaps. Even after the reduction the traces are still considerably long, with durations ranging from 3 days up to 25 days.

The SCADA traces were collected at two different water treatment and distribution facilities. The networks at these locations can be divided into two logical subnetworks. The lower level field network consists of programmable logic controllers (PLCs) and field devices and the higher level control network consists of servers with different functions (e.g.,

polling of PLCs, keeping historical data and performing access control), and Human Machine Interfaces (HMI), i.e., operator workstations. All communication between the networks has to traverse a gateway. In one of the locations this logical separation is also physical, i.e., there is no direct link between nodes in the field and control network. To capture all data from the second location, two separated measurements were done, one in each subnetwork. We treat these measurements as two different datasets, to which we refer as *s02t01* (control) and *s02t02* (field). A summary of the description of the datasets can be found in table I.

Since the goal of the paper is to characterize only the traffic specific to the SCADA system, we removed the traffic of the following network services by transport port numbers: 53 (DNS), 67-68 (DHCP),123 (Network Time Protocol), 137-139 (NetBIOS), 161-162 (SNMP), 546-547 (DHCP for IPv6). During the measurements, both locations were performing tests with IPv6 traffic that we also removed from the data.

IV. Analysis

Our analysis consists of three different tests that provide a high level characterization of the datasets. The time series analysis is based on previous research work done in [9]. Due to the increasing interest in Intrusion Detection Systems (IDS) for SCADA [4], [5], we consider time series analysis to be of particular importance, as it is widely used in IDS developed for traditional IT networks and the Internet [12].

In the two other tests we verify the validity of two expected characteristics of SCADA and SNMP. The first is the concept that these generate traffic in a highly periodical way, as a consequence of the polling mechanism used to retrieve data. The second is the assumption that they have a very stable connectivity graph, as nodes are not expected to be added or removed from the network. A number of other tests could be performed, but we argue that the analysis presented in this paper provides a general overview of important traffic patterns.

A. Periodicity

To check for periodicity, we carry out a Fourier analysis for the packet time series for each source IP address and the aggregate of all sources. We use 15s bins, thus the minimum period we are able to observe is 30s. Before applying a Fast Fourier Transform (FFT), we subtract the mean value of the time series from each entry in order to reduce the DC component. Zero-padding is used to adjust the length of the time series to a power of two. We artificially limit the maximum displayed period to 1800 seconds, as periodicity is not expected at larger time scales. Finally, e restrict our analysis to active sources, i.e., hosts that send at least 1 packet for at least 2% of the bins.

As expected, both SNMP and SCADA traces exhibit periodical behavior, although at different scales. Most of the SNMP traces show a periodicity at 300s, consistent with the default polling interval of many applications. The typical SNMP behavior is exemplified in Figure 1a, where the periodogram for the dataset *l15t04* is shown, presenting high power at the 300s component. The only exception in the SNMP datasets is *l05t01*, with a period of 60s.

The SCADA datasets exhibit periodicity at a smaller scale, with the datasets *s01t01* (Figure 1b) and *s02t01* presenting the dominant period component at 60s. There is also indication of periodicity (i.e., smaller peaks) at other frequencies. The third SCADA dataset *s02t02* also shows periodicity at 60s, however the dominant period component is at 50s.

Some nodes at the SCADA dataset *s01t01* do not present periodicity at the expected time scales. Figure 1c shows a periodogram constructed for one of the sensors in this dataset. There is no indication of periodicity in the range $[30-1800]$ seconds. To further investigate the issue, we used a much higher sampling frequency of 100 hertz to check for periodicity at smaller time scales. The sensors *do* generate data periodically, but at a smaller period of 1 second (note that Figure 1c features a wider x-axis). The other exception is one of the PLCs at the same location, which presents an unexpected dominant period component of 2h.

A summary of the periodicity analysis is presented in Table II. The table reports the main period of the aggregated traffic, the number of hosts that generate traffic in periodic and non-periodic fashion and the number of hosts for which the analysis was inconclusive. Not only the aggregated traffic is periodic, but the majority of source addresses generate traffic in a periodical way.

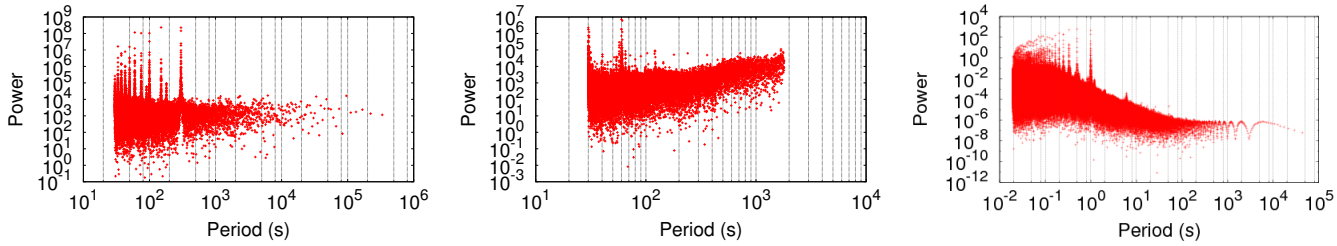
There are two exceptions in the SCADA traces, both in *s02t01*. One is a server that performs access control and the other a operator workstation. In both hosts, traffic patterns are dictated by human interaction, thus are not expected to be periodic. The unexpected behavior is that the other operator workstations *do* generate periodic traffic, probably related to some automated function (e.g, providing real-time information on the infrastructure). The majority of those workstations present a strong 60s period component. The last non-periodical source is an SNMP agent in the *l01t02* trace. The reason for this behavior is unknown.

trace	description	duration
l01t02	national research network	3.1 days
l05t01	regional network provider	15.4 days
l11t01	networking laboratory network	25.6 days
l15t04	national reserach network	7.7 days
l17t01	university network	5.0 days
l18t01	national reserach network	10.0 days
s01t01	water treatment facility	13.6 days
s02t01	water treatment facility (control)	10.2 days
s02t02	water treatment facility (field)	10.2 days

TABLE I: Overview of the datasets

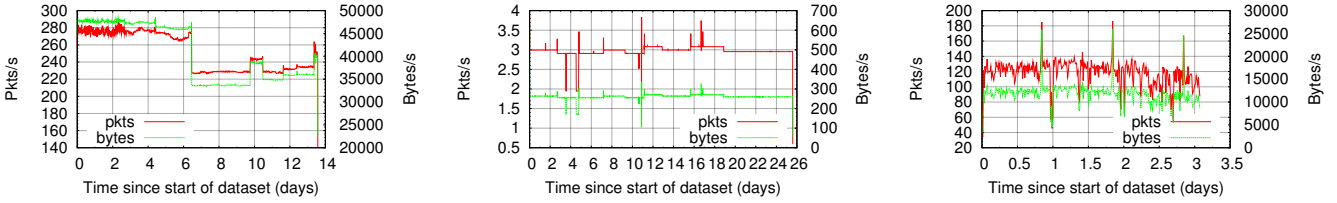
dataset	main period	#periodic	#non-periodic	#inconclusive
l01t02	300	171	1	5
l05t01	60	52	0	10
l11t01	300	11	0	11
l15t04	300	42	0	90
l17t01	300	19	0	16
l18t01	300	61	0	22
s01t01	60	16	0	1
s02t01	60	12	2	5
s02t02	50	8	0	6

TABLE II: Summary of the periodicity analysis



(a) l15t04 all (b) s01t01 all (c) s01t01 sensor

Fig. 1: Peridiograms



(a) s01t01 (b) l11t01 (c) l01t02

Fig. 2: Time series

B. Time series

All SCADA time series show clear constant throughputs over long periods of time, to which we refer as *baselines*. A notable feature is the change of the baseline, for example in Figure 2a between day 6 and 7. The main cause for this phenomenon is the start (or end) of high throughput flows. This happens at seemingly arbitrary times, with no apparent pattern. Due to the time scale in which it occurs, generally measured in days, and the high amount of changes we observe it is unlikely that these changes are caused by hardware/software failures. We speculate they are due to changes in the infrastructure, such as water tanks becoming full and pipes being closed.

In addition to the changes of the baseline, we also observe smaller variations and peaks of sudden activity. They are caused by momentary increase/decrease in the amount of variables requested by a monitor and/or in the rate in which the variables are requested. We assume that the peaks are caused by human-generated traffic. We know that some operator activities, such as uploading new configuration to the PLCs, can drastically increase traffic for a short moment.

The SNMP traces present, in general, a similar behavior, as shown in Figure 2b for the *l11t01* dataset. However, we note that the SCADA traffic exhibits more drastic baseline changes. One of the possible explanations is the small amount of hosts in the SCADA environments. When one host (or even one flow) starts/stops sending data, the relative impact on the total amount of traffic is larger.

One of the exceptions is shown in Figure 2c. The time series of the *l01t02* dataset shows a large amount of variance, but also some regular peaks (discussed in Section IV-C). The high number of consecutive packets requesting the same variable indicate packet loss, which could explain the observed variance.

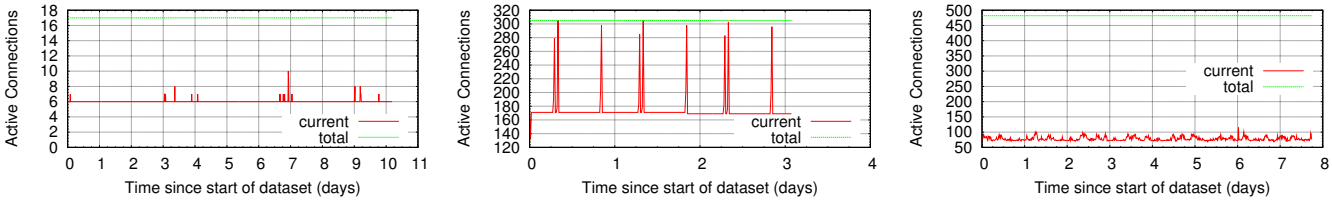
Finally, we point out that none of our datasets, SCADA nor SNMP, presents diurnal patterns often observed in network measurements [13]. Clearly, human activity does not have a major impact on the analyzed datasets.

C. Topology Changes

For our first test, we build the time series of the total number of connections using 15 minutes bins, as in Section IV-B. Figure 3 shows the results for three datasets. All SCADA datasets present a remarkably stable number of connections, with a clear baseline that keeps almost unchanged for periods longer than a day, as it can be seen in the results for *s02t02* presented in Figure 3a. The largest instability is a peak with 4 connections above the baseline.

Regarding this test, the SNMP datasets can be divided in two groups. *l01t02*, *l05t01* and *l11t01* have a behavior closer to that of SCADA, while the remaining datasets present more variation. An interesting result is shown in Figure 3b. Apart from the clear baseline, *l01t02* present very periodic peaks. After closer inspection, we verified that one of the monitors is only active during three moments, around 11h, 22h and 23h every day, causing the observed peaks. The result for dataset *l15t04* is shown in Figure 3c, as an example of the later group. It presents a baseline of approximately 93 connections, but with high variation in the range $[73 - 145]$.

In the second test we study how changes occur in the connection graph. First, we build a list with the active connections in each bin. We then compare consecutive bins, creating two lists: *new* and *missing* connections, i.e., connections present in the later bin, but not in the former and vice versa. For this analysis we use 1 hour bins, as our objective is to study



(a) s02t02 (b) l01t02 (c) l15t04

Fig. 3: Current IP connections

the long term stability the connection graphs. The results are summarized in Table III.

In average, the connection graphs do not present many changes over time, ranging from extremely low 0.1% up to 6%. The standard deviation, shown at table as *std*, is also low. Only three datasets present maximum changes of 20% or above. In *l01t02*, the high standard deviation and maximum can be explained by the periodical peaks mentioned before. The dataset *l11t01* presents a large peak in the number of connections close to the end of the trace. However, this is does not translate in a significant increase in traffic (bytes or packets). This high maximum is probably caused by some unusual behavior, such testing a new monitor configuration. The high maximum in the SCADA dataset *s02t02* can be partially explained by the low amount of hosts in the network. In this dataset, 29% represents an increase of 4 connections.

In spite of the small number of changes in each bin, the occurrence of changes is the rule, rather then the exception. This can be observed in the column *changed*, where the percentage of bins that contain at least one change is reported. For most of the cases, the majority of bins contain changes. This means that even for the cases where the number of active connections is stable, the connectivity graph continuously changes. These observations need to be incorporated in models that describe SCADA operation. For instance, the IDS proposed in [5] generates alarms for each new observed flow, reporting it as an anomaly. If this IDS was to be deployed in one of our SCADA datasets, a large amount of (false) alarms would be generated.

V. Conclusions

This paper provides a first look into the characteristics of SCADA network traffic. Our results show that, to some extent, SCADA traffic is similar to SNMP traffic. Both present remarkably regular time series, due to the fact that the majority of the sources generate data in a periodical fashion. However, one should observe that changes do occur. Time series present baseline changes at seemingly arbitrary time intervals. Some of the hosts do not generate traffic in a periodical fashion. Finally, although small, the occurrence of changes in the topology is the rule rather then the exception.

Our initial analysis showed some interesting results, however more research is needed to fully understand the particularities of SCADA traffic. Other characteristics should be studied and it is necessary to verify if our results also apply to other SCADA networks. In future work, we intend to extend the analysis presented here, with the goal of providing an empirical foundation for future research in the area.

dataset	new			missing			changed
	mean	std	max	mean	std	max	
l01t02	4 %	11%	42%	4%	11%	44%	98.63%
l05t01	0.4%	1%	18%	0.4%	1%	17%	32.35%
l11t01	0.1%	2%	33%	0.1%	1%	29%	3.9%
l15t04	4 %	4%	15%	4 %	4%	15%	100%
l17t01	4 %	3%	17%	4 %	3%	14%	95.80%
l18t01	3 %	2%	14%	3 %	2%	14%	99.17%
s01t01	4 %	2%	9%	4 %	2%	12%	93.23%
s02t01	3 %	3%	19%	3 %	2%	16%	79.92%
s02t02	5 %	4%	29%	6 %	4%	29%	84.84%

TABLE III: Topology changes

References

[1] J. Case, M. Fedor, M. Schoffstall, and J. Davin, "RFC 1157: Simple network management protocol (SNMP)," *IETF, April*, 1990.

[2] R. Barbosa, R. Sadre, and A. Pras, "Difficulties in modeling scada traffic: A comparative analysis," in *Passive and Active Measurement*. Springer, 2012.

[3] V. Igure, S. Laughter, and R. Williams, "Security issues in SCADA networks," *Computers & Security*, vol. 25, no. 7, pp. 498–506, 2006.

[4] S. Cheung, K. Skinner, B. Dutertre, M. Fong, U. Lindqvist, and A. Valdes, "Using model-based intrusion detection for SCADA networks," in *Proceedings of the SCADA Security Scientific Symposium*. Citeseer, 2007, pp. 1–12.

[5] A. Valdes and S. Cheung, "Communication pattern anomaly detection in process control systems," in *Technologies for Homeland Security, 2009. HST '09. IEEE Conference on*, IEEE. IEEE, May 2009, pp. 22–29.

[6] W. Stallings, *SNMP, SNMPv2, SNMPv3, and RMON 1 and 2*. Addison-Wesley Professional, 1999.

[7] L. Andrey, O. Festor, A. Lahmadi, A. Pras, and J. Schonwalder, "Survey of SNMP performance analysis studies," *International Journal of Network Management*, vol. 19, no. 6, pp. 527–548, Nov. 2009.

[8] T. Oetiker, "MRTG - The Multi Router Traffic Grapher," in *Proceedings of the 12th USENIX conference on System administration*. Berkeley, CA, USA: USENIX Association, 1998, pp. 141–148.

[9] J. Schonwalder, A. Pras, M. Harvan, J. Schippers, and R. van de Meent, "SNMP Traffic Analysis: Approaches, Tools, and First Results," *2007 10th IFIP/IEEE International Symposium on Integrated Network Management*, pp. 323–332, May 2007.

[10] G. van den Broek, S. ten Hoeve, G. C. Moreira Moura, and A. Pras, "SNMP Trace Analysis: Results of Extra Traces," Centre for Telematics and Information Technology University of Twente, Enschede, Technical Report TR-CTIT-10-06, Dec. 2009.

[11] "TCPDUMP." [Online]. Available: http://www.tcpdump.org/

[12] A. Sperotto, G. Schaffrath, R. Sadre, C. Morariu, A. Pras, and B. Stiller, "An overview of ip flow-based intrusion detection," *Communications Surveys & Tutorials, IEEE*, vol. 12, no. 3, pp. 343–356, 2010.

[13] S. Floyd and V. Paxson, "Difficulties in simulating the Internet," *IEEE/ACM Transactions on Networking*, vol. 9, no. 4, pp. 392–403, 2001.