URSI GASS 2020, Rome, Italy, 29 August - 5 September 2020

A Performance Study of the Latent Dirichlet Allocation Technique Applied to Sequence Identification in 802.15.4 Data Link Layer Traces

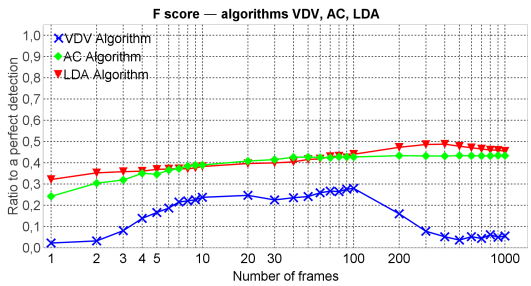
Pierre-Samuel Gréau-Hamard*[(1)(2)], Moïse Djoko-Kouam[(1)], and Yves Louet[(2)]
(1) Informatics and Telecommunications Laboratory,
ECAM Rennes Louis de Broglie, Bruz, France,
http://www.ecam-rennes.fr
(2) Signal, Communication, and Embedded Electronics (SCEE) team,
Institute of Electronic and Telecommunications of Rennes (IETR), CentraleSupélec, Rennes, France,
http://www-scee.rennes.supelec.fr

We place ourselves in the context of a communicating object coming into an unknown environment and wanting to establish a communication with the existing networks. To that end, we aim to learn the unknown protocols of the environment, and not just identify them from a database.

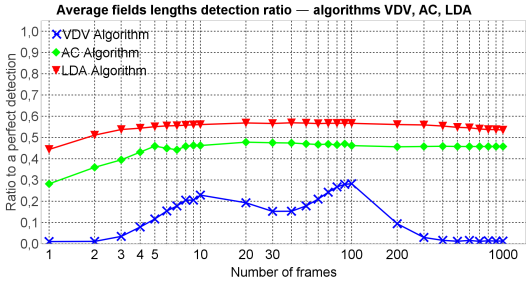
This is the goal of Protocol Reverse Engineering (PRE), a family of techniques which aims at reconstructing the frame formats and/or the state machine of a target unknown protocol through analyzing execution traces and/or network traces. The general procedure to perform PRE, is a five-step process. Firstly, (i) the radio traffic is intercepted and the frames issued by the targeted protocol are isolated. Next, (ii) the meaningful binary sequences (features) of these frames are identified, (iii) then the frames are grouped by format via the use of these features. Within each group, (iv) sequences alignment is performed, and, finally, (v) the frame formats and/or the state machine of the targeted protocol are reconstructed. Here, we focus solely on the second step, the identification of remarkable sequences. This step aims at reducing the quantity of information needed to label a frame.

In a previous paper [1], we compared the performance of three techniques achieving this : Variance of the Distribution of Variances (VDV), Aho-Corasick (AC), and Latent Dirichlet Allocation (LDA)[2]. To perform this comparison, we simulated the three algorithms associated with randomly generated 802.15.4 Data Link Layer (DLL) traces (from Zigbee stack), and we found that LDA was by far the best performing one. So, in this paper, we will compare the performance of LDA obtained in [1] with that of a trace sniffed on a full-blown network, containing real communications.

As comparison metrics, we will use **precision** and **recall** to quantify the relevance of the sequences detected, as well as a tradeoff between the two, the F score. We also use our own metric, the **fields detection ratio**, to quantify the detected information of the fields. We hope to present results as good as the ones in figures 1a and 1b, obtained in [1].



(a) Best F score of the algorithms VDV, AC, and LDA



(b) Best average field lengths detection ratio of the algorithms VDV, AC, and LDA

References

[1] P.-S. Gréau-Hamard, M. Djoko-Kouam, and Y. Louet, "A comparative study of sequence identification algorithms in iot context," accepted to ASPAI' 2020, Berlin, April 2020.

[2] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, May 2003.