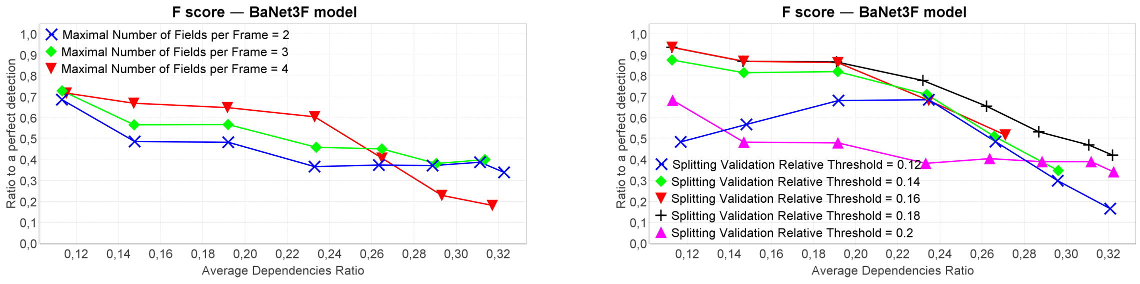
A Bayesian Networks based Automatic Binary Frame Format Identification Model : BaNet3F

Pierre-Samuel Gréau-Hamard*[(1)(2)], Moïse Djoko-Kouam[(1)(2)], 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,
Institut d'Electronique et des Technologies du numéRique (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 aiming to establish a communication with the existing networks. To that end, the object needs to be able to adapt itself to whichever standard used in the targeted environment. It is the same goal as the one pursued by Software Defined Radio, except that in our case, we propose to learn the unknown protocol of the environment, and not just identify it from a database.

So, in this work we developped a model designed to reconstruct the frame formats of a target unknown binary protocol through analyzing its network traces. This model, called Bayesian Network Frame Format Finder (BaNet3F), is based on the Bayesian networks theory. It aims to learn the parameters of a Bayesian network used as a generative model of the frames to analyze. Once the learning procedure is over, the most probable state of the network is calculated for each frame, and its hidden structure is retrieved. We compared the performances on synthetic traces of BaNet3F with two models of the state of the art : LDA [1] and Cai et al [2], and our model was always better. However in this paper we will focus on evaluating the performances of the model on synthetic and real binary data link traces and comparing the two. The real traces will follow the IEEE 802.15.4 standard, while the synthetic ones will be entirely created from scratch, without following any standard.

As performance metrics, we use the precision and recall to quantify the relevance of the fields identified, as well as the F score, the harmonic mean of precision and recall, to give us a tradeoff. We will also use a complementary metric as a reference to evaluate the complexity of the traces analyzed : the Average Dependencies Ratio (ADR). It quantifies the ratio of the dependencies between fields to the dependencies within fields. The higher ADR is, the more difficult the trace to be analyzed by BaNet3F. Figures 1a and 1b show some of the performance graphics we obtained through simulating BaNet3F applied to synthetic traces.



(a) F score of the BaNet3F model, as function of the ADR, parameterized by the maximal number of fields per frame, and averaged over 5 points

(b) F score of the BaNet3F model, as function of the ADR, parameterized by the splitting validation relative threshold, and averaged over 5 points

References

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

[2] J. Cai, J.-Z. Luo, and F. Lei, "Analyzing network protocols of application layer using hidden semi-markov model," *Mathematical Problems in Engineering*, vol. 2016, pp. 1–14, 01 2016.