PROPOSAL for
Interdisciplinary project in Machine Learning and Data Analysis
Sep 2020, FIV

Student:
Allan Kálnay

Co-Supervisor:
Félix

Master/Bachelor/Comp.project:
Inter. project in ML and DA

Topic (provisional title):
TCP/IP communication flows into sentence-like transcriptions

Project number:
SPT.4

Research Question:
Design a suitable symbolic schema that transforms TCP/IP flows into sentence-like transcriptions and evaluate its suitability for knowledge discovery and (ideally) attack detection.

Pre-requisites:
Programming skills in Python
Programming skills in Perl
Basic knowledge in TCP/IP protocols and communications
Basic knowledge in statistics and data analysis
Basic knowledge in supervised and unsupervised classification

Methodology:

1. Get familiar with the field and context of the project. Check out the provided code (pcap2transcription) and the related literature.
2. Get familiar with network traffic data for testing: (a) the CIC2017 IDS dataset (synthetic, labeled), and MAWI datasets (real, unlabeled).
   CIC2017: https://www.unb.ca/cic/datasets/ids-2017.html
   MAWI WIDE: https://mawi.wide.ad.jp/mawi/
3. Propose a map of key symbols to transform network traffic data into text-like flow transcriptions that are suitable for encrypted communications. Let‘s preliminarily call it NTFT (from Network Traffic Flows Transcriptions).
4. Develop scripts and toy examples for testing, proof of concept, and knowledge extraction.

Expected Outcome:

- A description of the ntft symbolic language.
- A set of processing scripts for transforming pcaps into ntfts (python)
- A set of scripts for analyzing ntfts files (python)
- Ntft-based description of network traffic data (CIC2017, MAWI, other??)
- Toy examples that work out-of-the-box (python)
- Project report

Expected Software:
Python

Notes and tips:
You might optionally want to check bag-of-words, MinHash-based classification, or LSH (local sensitivity hashing). If so, ask me for clarifications and example scripts.

Deadlines and milestones:

- Next meeting to propose by student after completing the step 1 of the Methodology.

Related literature:

[1] Meghdouri, F.; Zseby, T.; Iglesias, F. Analysis of Lightweight Feature Vectors for Attack Detection in Network Traffic. *Appl. Sci.* **2018**, *8*, 2196.

[2] Novo, C., & Morla, R. (2020). Flow-based detection and proxy-based evasion of encrypted malware C2 traffic. *arXiv preprint arXiv:2009.01122*.

[3] Kakavand, M., Mustapha, N., Mustapha, A., & Abdullah, M. T. (2015). A Text Mining-Based Anomaly Detection Model in Network Security. *Global Journal of Computer Science and Technology*.

[4] Suh-Lee, C., Jo, J. Y., & Kim, Y. (2016, October). Text mining for security threat detection discovering hidden information in unstructured log messages. In *2016 IEEE Conference on Communications and Network Security (CNS)* (pp. 252-260). IEEE.

[5] Dionísio, N., Alves, F., Ferreira, P. M., & Bessani, A. (2019, July). Cyberthreat detection from twitter using deep neural networks. In 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[6] Xie, G., Xie, K., Huang, J., Wang, X., Chen, Y., & Wen, J. (2017, May). Fast low-rank matrix approximation with locality sensitive hashing for quick anomaly detection. In IEEE INFOCOM 2017-IEEE Conference on Computer Communications (pp. 1-9). IEEE.

[7] Sael, L., Jeon, I., & Kang, U. (2015). Scalable tensor mining. Big Data Research, 2(2), 82-86.