Brufence, WP2: Communication Systems Security

SPICES and BRUFENCE Joint User Committee meeting
22 November 2016

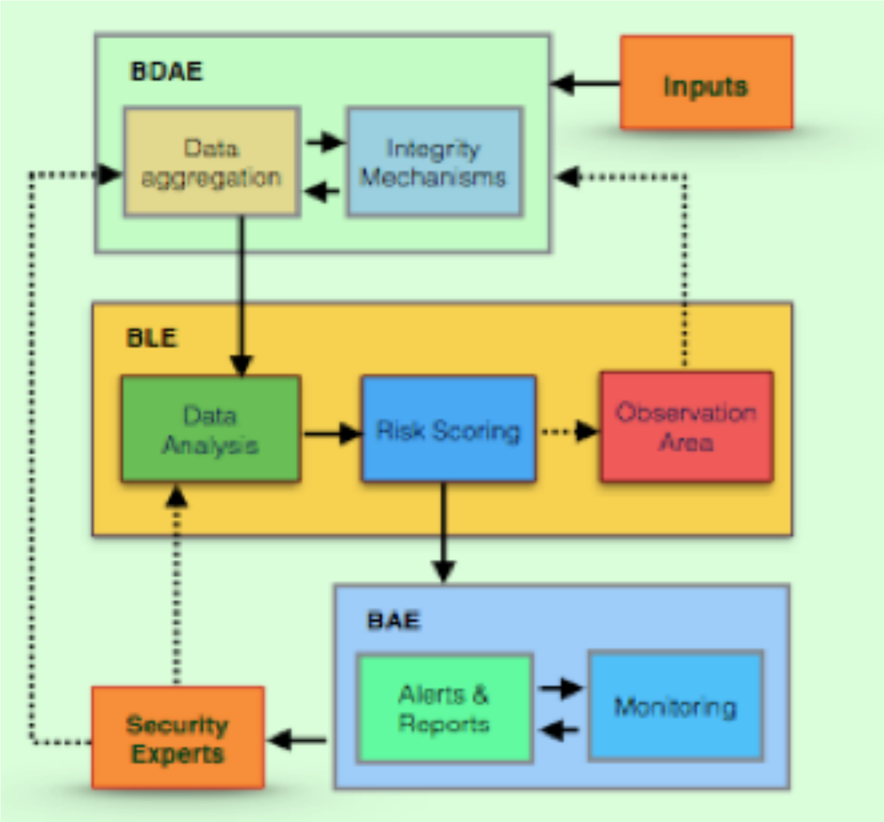
Dimitrios Sisiaridis - Olivier Markowitch

QualSec Group Departement d'Informatique, ULB

Q4-Q6 (1/2016 - 9/2016): Design of Analysis Tools (-> **Q10:** 9/2017)

		Y1				Y2				Y3			
		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
WP 2	Communication Systems Security	X	X	X	X	X	X						
T 2.1	Market Intelligence	X											
T 2.2	Analysis of existing T&P		X										
T 2.3	Project Procurement			X									
T 2.4	Design of Analysis Tools				X	X	X						
T 2.5	Interfaces Design												

The Framework



On-going tasks I

- *Data Acquisition module* (AM)
    - focused on interoperability, data anonymization and automated procedures
- *Exploratory Data Analysis sub-module* (part of the *Data Analysis module* of the Learning Engine - LEM)
    - using the python Spark API, releases 1.6.2, 2.0.1
    - using/extending external python libraries (scikit-learn, pandas, nltk)
- *Predictive Analytics* in the LEM
    - machine learning algorithms of Spark core libraries (MLlib, spark streaming) and HiveSQL (on premises)
    - AWS EMR and EC2 (on Amazon cloud)
        - elasticity on AWS facilitates a more scalable solution against on-premises big data analytics

On-going tasks II

- *Alert and Report Module* (ARM)
  - based on building a recommendation engine
- the *Active Learning sub-module*, as a connector between the experts and the Data Analysis sub-module
  - currently by using ‘simulated’ feedback, to improve the accuracy of the classifier following clustering of un-labelled data
- *Causal ordering* techniques for handling missing steps in the kill chain
  - development of novel constructions for reproducing missing links in threat and attack patterns utilising category theory and functional logic
    - to study potential attacks based on risk management by studying the behaviour of users, services and network
    - to re-construct or synthesise the remaining phases of the kill-chain, in order to understand adversary’s intents, techniques and tactics.

Time-series analysis on temporal and spatiotemporal data I

- Activities are decomposed in time intervals on a basis of (sec), minute, hour, day, week, month
  - to maintain aggregations that can be used to satisfy real-time requirements
  - in order to reduce the computational effort and ultimately the computational cost
- in spatiotemporal data:
  - trajectories are measured over time
    - to detect attacker’s intentions in kill-chains
    - to trace-back zero-day attacks, based on re-usable patterns, to earlier phases of the intrusion kill-chain

Time-series analysis on temporal and spatiotemporal data II

- Spatial data:
  - non-spatial data are measured at spatial locations
  - unusual changes are reported as outliers (substantial different behaviour from those in their neighbourhood)
  - contextual attributes define the location of interest
  - behavioural attributes are measured for each object
- Non-stationary data (e.g. in order to detect event detection in texts):
  - changes that differ significantly from the trend (e.g. first story detection)
    - trend-stationary series, by fitting a trend-line including the time-index
    - difference-stationary series, by transforming the series over periodical differences

Models employed

- *Clustering* for un-labelled datasets (un-supervised analysis)
  - variations of the k-means clustering algorithm (k-means, k-means soft, bisecting, gaussian clustering)
- *Classification/Regression* for labelled and partially labelled datasets (supervised & semi-supervised analysis)
  - linear/logistic regression, Support Vector Machines (SVM), Naive Bayes, decision trees, random Forest, ensemble methods
  - active learning, based on a recommendation engine: to improve the accuracy of the classifier by allowing to choose a subset of the learning data
  - a *voting scheme* for the ensemble models metric values
- we examine the use of neural networks for implementing novel categorical structures to control system's state through HMM and linearizations

Transformation methods

- Text mining methods for pattern matching:
  - **Word$_2$Vec:** to compute distributed vector representation of words in websites and texts
  - **Term Frequency - Inverse Document Frequency** (TF-IDF):
    - *Term Frequency* (TF), i.e. the importance of a term, that is the number of times the term appears in a document
    - *Inverse Document Frequency* (IDF): looking to minimise the number of times a term appears in all documents/texts under analysis
- regularisation for linear problems
  - **StandardScaler:** in SVM and linear regression
- text classification and clustering
  - **Normaliser:** based on the cosine similarity to transform input vectors
- categorical features representation:
  - transformation to dictionaries and then using either:
    - binary representation using sparse vectors
    - the probability distribution of key entries
    - the actual key entries, followed by a lookup reference to the actual values

Feature selection and collaboration filtering methods

- Feature selection:
  - **ChiSqSelector** (currently only for the public datasets)
  - **PCA:** in multi-dimensional spatiotemporal time series, to project vectors in a low-dimensional space
  - the **feature importance** metric value: using the randomForest algorithm for feature correlation to the output value
- Collaboration filtering:
  - **Alternating Least Squares** (ALS): in the recommendation engine of the alert module

Evaluation metrics

- Classification metrics (linear):
  - **Mean Squared Error** (MSE)
  - **Mean Absolute Error** (MAE)
  - **Mean Squared Log Error** (MSLE)
  - **R-squared** (R^2) (how close the data are to the fitted regression)
  - **Explained Variance** (ER)
- classification & logistic regression metrics:
  - **Sensitivity** (i.e. **Recall**): TruePositiveRate = TruePositives / (TruePositives + FalseNegatives)
  - **Specificity** (i.e. **Precision**): TrueNegativeRate = TrueNegatives / (TrueNegatives + FalsePositives)
  - **Accuracy:** the number of correctly classified examples / total examples
  - **Precision-Recall** (P-R) Curve
  - **ROC** curve: TruePositiveRate against FalsePositiveRate
  - **Area Under the ROC** curve (AUC)
- clustering metrics:
  - **Intra-Cluster Distance** and **High-Silhouette Coefficient**
    - by examining *similarities*, *variance* and *outliers* which do not correspond to security noise

Performance measurements

best known metrics using a similar approach (MIT/PatternEx)

- improve attack detection rates (by having 200 events daily labeled by the analysts)
  - detection rate: 86,8%
  - false positive rate: 4,4%
- reduce security noise
  - by reducing the number of alerts to the security analysts

- our results:
  - *best accuracy* (with scaled data): 0.722837878
  - *Area Under Presicion-Recall* (APR): 0.790599748 (using decision trees with the impurity parameter set to ‘gini’)
  - *Area Under the ROC* curve (AUC): 0.699113503
    - which indicates the need for further research on tuning and optimisation of the employed algorithms, possibly via ensemble methods

Valorisation activities

- in **Q5-Q6**:
  - Conference papers:
    - Sisiaridis D., Carcillo F., Markowitch O., A Framework for Threat Detection in Communication Systems, PCI 2016, 20th ACM Conference on Informatics, 10-12 November 2016, Patras, Greece
    - Sisiaridis D.,Kuchta V.,Markowitch O.,*A Categorical Approach in Handling Event-Ordering in Distributed Systems*, ICPADS 2016, 22nd IEEE International Conference on Parallel and Distributed Systems, 13-16 December 2016, Wuhan, China
    - we are currently working on a paper submission to the:
      - *IEA/AIE NAAD 2017 : Special Track on Novel Approaches to Anomaly Detection, 30th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 27-30 July 2017, Arras, France
      - another paper on *causal ordering and its implication to the kill-chain model* in due time before the end of Q8
  - a series a meetings and seminars with the industry
    - Sopra Steria & Sopra Steria Banking
    - the Institut Belge des services Postaux et des Télécommunications (IBPT)
    - ING Bank
    - SNA - Brussels Airlines

Next steps

- evaluate the models against real use cases (WP5.2: Case Study II - Implementation of the strategies developed in WP2, **Q7-Q12**):
  - bank sector, telecommunications, airline companies, public sector
- strengthen the algorithms e.g. against data contamination (**Q8-Q10**)
  - by using public blockchains
  - by tuning model parameters
  - by verification protocol analysis
- continuous model refinement
- automate processes in all modules of the engine (**Q8-Q11**)
- employing *sandboxing* and implement the novel categorical structures for *causal ordering* (for further analysis of adversary's intents, techniques and tactics) in the relevant stages of the kill chain (**Q7-Q10**)
- design interfaces for real-time interaction (**Q9-Q12**)
  - work on *threat intelligence*: e.g. communicate findings in real-time through the distributed nodes of a network

Thank you

olivier.markowitch@ulb.ac.be, dimitrios.sisiaridis@ulb.ac.be