Brufence Project - Work Package #2: Communication Systems Security Detection of Threats and Attacks on Managed File Transfer and Collaboration Platforms

Olivier Markowitch
Department of Computer Science
QualSec Group—Universit*é* Libre de Bruxelles
Faculty of Sciences, Campus de la Plaine
ULB CP212, Boulevard du Triomphe
1050, Brussels

Dimitrios Sisiaridis
Department of Computer Science
QualSec Group—Universit*é* Libre de Bruxelles
Faculty of Sciences, Campus de la Plaine
ULB CP212, Boulevard du Triomphe
1050, Brussels

23 October 2015

Abstract

This report is a summary of the general status, research progress, difficulties and next stages of the work package #2 of the Brufence project, on scalable machine learning for automatic detection of threats and attacks in communication systems.

1 Introduction

Managed File Transfer software is a class of integration middleware used by enterprises as well as by public and private organizations for secure and guaranteed delivery of a file or set of files from a source to a target over a network e.g. from an organization to another directly or via a file transfer service provider or within the organization through a collaboration platform irrespectively of users? location, in order to improve operational efficiency through the automation of system-centric activities. Examples of commercial products are the AxWay MFT, IBM MFTs suite (WebSphere, Sterling, Aspera), Oracle MFT, Ipswitch MOVEit MFT, Attunity Gateway, MIX MFT, SIS HULFT, Primeur Spazio as well as open source solutions such as the SOS MFT, BIT MFT, JADE or the DivConq MFT.

Collaboration platforms & tools are a combination of collaborative software supporting communication, conferencing and coordination activities in enterprises as well as in organizations in public and private sector. Examples of commercial products are MS Sharepoint, Alfresco, the eXo Enterprise Social Platform, Huddle, Abiquo True Hybrid Cloud, Zimbra Collaboration, Liferay, IBM Connections v.3, ProjectLibre, Clarion, Wrike, SpiraTeam, GoToMeeting, BaseCamp, WebEx Meeting Centre, Bitrix24, Redbooth, Tamashare, SAP Stxeamwork, SocialText, Zoho Apps and Google Apps as well as the OwnCloud open source platform.

Work Package #2 (WP2) of the Brufence project aims to the design of a near real-time scalable modular framework for deploying machine learning (ML) algorithms for Big Data predictive analytics in the field of threat detection in collaboration platforms and Managed File Transfer (MFT) systems in terms of a prototype system. Different models will be compared in parallel in order to maximise their efficacy in predicting abnormal behaviour, evaluated for their performance in terms of scalability, accuracy, readability and QoS.

The system will be in compliance with the CERT-EU guidelines for threat detection. It is intended to be a multi-layer, modular approach with mechanisms covering the set of actions proposed by the EC Guide for identification of threats and attacks, such as detecting and alerting, attack analysis, motivation identification as well as threat mitigation and refinement. Furthermore, it promotes a centralised monitoring, processing and analysis of complex events in MFTs and collaboration platforms reducing thus the impact of the visibility of data movement inside and outside (i.e., data exfiltration) the enterprise or the organisation that deploys collaboration software.

2 General Status

Progress in WP2 is in line with the schedule in terms of time and deliverables. An analysis of the state-of-the-art methodologies on the literature as well as a survey of the best practices in commercial products and open source solutions in the field of auto-protected systems in terms of threat and attack detection by deploying machine learning algorithms are already completed (*WP2.1: Analysis of existing techniques and products*, *WP2.2: Market Intelligence*), including the exchange of information with the major vendors in the field.

3 Research Outcome

In the Brufence system, automated traffic log analysis of temporal and spatial data, over a long period of time, will be engineered for advanced proactive threat protection. Real time log data acquired by network devices in a centralised log management system following e.g. an SDN approach, will be correlated with attack communication profiles, derived from a learning set of identified baseline behaviour profiles of indicators of compromise, as well as from multiple intrusion kill chains over time, representing a complete picture of how an adversary acts in a variety of environments aiming to achieve a rapid and accurate identification of threat patterns in order to detect or even to predict Advanced Persistent Threats (APTs), DDosS attacks and zero-day attacks.

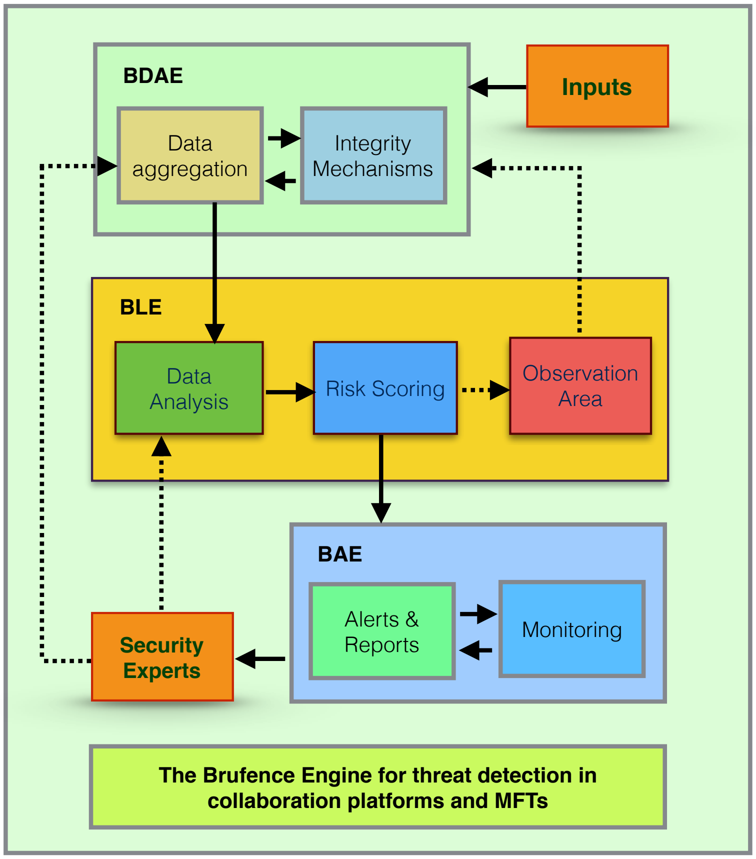
Un-structured, semi-structured and structured data with respect to security -related events from users, services and the underlying network infrastructure, with a high level of large dimensionality and non-stationarity as well as temporal aspects, will be correlated and classified, in order to detect abnormal behaviours that deviate from normal behaviour. Incident correlation integrated with complex event processing techniques will be used to compare different events, often from multiple data sources, in order to identify patterns and relationships which enable the identification of events belonging to an attack or, are indicative of a broader malicious activity. In this way, they will allow a better understanding of the nature of an event, reduce the workload needed to handle incidents and automate the classification and forwarding of incidents only relevant to a particular consistency reducing thus security noise and false positives.

This processing and analysis of complex events will be materialised by deploying machine learning algorithms in terms of *outlier detection analysis* of threats and attacks in collaboration platforms and MFTs. The most significant challenge for an evaluation of threat and attack detection algorithms is the lack of any appropriate public relevant datasets, mainly due to restriction rules applied on sensitive data by organisations and enterprises regarding confidentiality as well as potential security breaches.

3.1 The Brufence Engine

By adapting or rewriting ML algorithms, the proposed framework aims to enhance system functionalities for the analysis of huge amount of log data aggregating by a *Data Acquisition Engine* (BDAE), particularly in cases where there is a large amount of security noise, large dimensionality and non-stationarity of data, as well as the need for rapid and accurate identifications of threat patterns. The proposed *Learning Engine* (BLE) will enable the consumption of seemingly unrelated disparate datasets, regardless of their format, to discover correlated patterns that result in consistent outcomes with respect to the access behaviour of users, network devices and applications involved in risky abnormal actions, and thus reducing the amount of security noise and false positives. Along with history- and user-related data, network log data will be exploited to identify abnormal behaviour concerning targeted attacks against the underlying network infrastructure as well as attack forms such as man-in-the-middle and DDoS attacks. Network connectivity and analysis of traffic data can provide valuable information to identify a single-point of attack and failure as well as the most critical or vulnerable between connected nodes, in order to improve the accuracy of the predictive model, based e.g. on graph-based classification or semi-supervised classification models.

Real operational logs taken from collaboration platforms and MFTs installed on premises or in the cloud such as MS Sharepoint, OwnCloud or other resources will be used for Big Data analytics, utilising outlier detection analysis algorithms. Hadoop ecosystem, including HDFS, HiveSQL, Impala, Cassandra, Hue, Spark MLib functionality with an API implementation in Scala, along with the adaptation of algorithms provided by the open source ELKI platform written in Java as well as the re-definement of other ML algorithms, especially in the case of unsupervised learning, as well as the use of Apache Flume or Kafka for data aggregation along with python scripts for log analysis and pattern matching are among the tools that will be engineered in the next stage. Spark Streaming will be used for real-time stream processing on ML workflows for predictive analytics.



In scenarios as those with the analysis of operational logs from a collaboration platform or an MFT with time series and multidimensional streaming, outlier detection of unusual events needs to be performed in a time-critical manner, also referred as *streaming outlier detection*. *Outliers* are defined in terms of an abrupt change detection in the time-series, which corresponds to sudden changes in the trends in the underlying data stream such as those in time-series values with respect to immediate history, or distinctive shapes of subsequences of the time series, with respect to long-term history. In such cases, there are relationships among the data points of the dataset set of values are generated by continuous measurement over time (temporal data). Time-series data that form complex events, derived by the underlying network, middleware and authorisation, authentication & auditing (AAA) services. In this case, outlier detection is highly related to the problem of *change detection* where normal models of data values are highly governed by adjacency. New data values are referred to as *novelties*. A change could happen either slowly over time (known as *concept drift*) or abruptly. In the first case, it can only be detected by a detailed analysis over a long period of time, where in the latter, a suspicion of a possible change to the underlying data generation mechanism, detected with real-time analysis, can be taken as an indication of an anomaly, in terms of an *Advanced Persistent Threat*. Semi-supervised novelties can be either, only instances of a subset of the normal and anomalous classes may be available, where some of the anomalous classes may be missing from the training data, or labeled examples of all normal and some anomalous classes are available, where labels for the anomalous classes are not exhaustive. In the case of *active learning*, human feedback is utilised in order to identify relevant outlier examples.

In traffic log analysis with data representing a sequence of accesses in an operational log, the goal is to determine the unusual patterns of accesses in this log. Operational logs are typically pre-processed into a set of user-specific discrete sequences which correspond to the identifiers of e.g. the pages or services accessed by the users. Users can often be distinguished only at the level of their IP-addresses while user sessions can often be mined from the logs; the log needs to be decomposed into user sessions and then, it is further decomposed into test sequences, and comparison units. In order to detect position outliers, the goal is to determine any unexpected accesses. In the case of contextual anomalies corresponding to a single unpredictable access, this can be accomplished by using either Markov or rule-based models. When the aim is to detect combination outliers, the goal is to determine unusual subsequences in the test sequence, by using either unsupervised methods such as window-based nearest neighbour models or Hidden Markov Models, or supervised methods such as the extraction of relevant features (e.g. k-grams).

In host-based intrusion detection at the nodes of the underlying infrastructure of a collaboration system on cloud or on premises, operating system call traces are available in the form of symbolic sequences; any anomalous subsequences in these traces correspond to malicious computer programs. In this case, data are similar to web logs at a conceptual level. Different kinds of applications that run on client-side for exchanging data with the platform execute different sequential combinations of calls. This sequential ordering of the calls provides critical information in order to distinguish between normal and malicious software. In the feature extraction phase, logs are transformed into symbolic sequences. In the case where commands are coming from multiple sources, they can be separated out into their different hostnames in order to examine the malicious behaviour of a particular host, usually by applying discrete sequence methods.

Having a stream of network packets or data records containing both continuous and categorical attributes, the goal is to detect in real or near-real time threats and attacks. Each individual record is multidimensional containing features extracted from the unit of network data (e.g., packet), or raw tcpdump data. Unsupervised learning algorithms will aim to detect aggregate change point in order to identify network-wide traffic anomalies corresponding to network intrusion and network attacks. Streaming and supervised novel-class detection methods will aim to identify repeating attacks in order to detect novel classes as new intrusions arise.

In the case of event detection in text files in order to detect e.g. data exfiltration of sensitive corporation data, the goal is to determine unusual documents, which differ significantly from the trend. In first story detection where a stream of documents is available, the goal is to determine unusual events corresponding to new topics in the stream of documents. User activities, may provide early knowledge of unusual events as changes; both supervised and unsupervised methods can be used, upon the availability of training data.

In the case of analysing information exchange e.g. by having a stream of emails, the goal is to determine the subset of emails which correspond to spam. Unsupervised methods can be used for unusual topic detection although results are often likely to be inaccurate. Working with supervised methods, specific features of the emails are learned which are then related to spam labels in the training data. The *Bayes classifier* or email spam filtering methods can be applied for text classification by taking into account specific email characteristics of a junk email such as the domain of the sender, any peculiar phrases or overemphasised punctuation or whether the recipient of the message was a particular user, or a mailing list.

Discrete sequences on the log files can be either temporal (categorical or time series data) or not temporal (based on their relative placement with respect to one another). In spatial data, non-spatial attributes are measured at spatial locations. Any unusual local changes in such values are reported as outliers. Spatiotemporal data are a generalisation of both spatial and temporal data. Their analysis can enlighten attacker′s intentions in *kill chains* and thus, helping the detection and tracing back of *zero-day* attacks. In network and graph data, data values correspond to nodes (structural data) while relationships among the data values correspond to the edges. Outliers may be modelled in different ways, depending upon any irregularity of nodes, in terms of their relationships to other nodes, or of edges.

4 Next Steps

In the next stage, security assessment of the basic building blocks will take place (*WP2.3: Project procurement*) as well as security analytics in terms of pattern matching and predictive analytics. Open source tools will be used extensively for the needs of this research regarding encryption, security assessment of the basic blocks, virtualisation, collaborative systems, managed file transfer, static/dynamic code analysis and traffic log analysis. Computational means for clustering and distributed/parallel computing in next steps, regarding Big data analytics and behavioural analysis, will be covered partially with the use of a public cloud (AWS/EC2), according to the schedule as well as the procurement of the relevant hardware equipment. Furthermore, adversaries behaviour will be studied with the use of hybrid-interaction server/client-side honeypots; sandboxes will be used for monitoring in parallel with client-side honeypots.