Flow Analysis for a Live APT Detection - *Data at Rest Analysis*

Abdullah Said AL-Aamri, Rawad Abdulghafor, Akram M Z M Khedher, Shuhaili Bt Talib

1st Abdullah Said AL-Aamri
PhD Student, IIUM
Kuala Lumpor, Malaysia
abdullah.alaamri@tie.com.om

2nd Asst. Prof. Dr. Rawad Abdulghafor
Asst. *Prof at IIUM, KICT*
Kuala Lumpor, Malaysia
rawad@iium.edu.my

3rd Prof. Dr. Akram M Z M Khedher
Prof at IIUM, KICT
Kuala Lumpor, Malaysia
akramzeki@iium.edu.my

4th Asst. Prof. Dr. Shuhaili Bt Talib
Asst. Prof *at IIUM, KICT*
Kuala Lumpor, Malaysia
shuhaili@iium.edu.my

Abstract- *Smartness of the new era, the era of technology, is set on an ad-hoc architecture, to deliver a variety of known and unknown services, where the difficulty of achieving sustainable, secure, and trustworthy data transmission is a big challenge. Even within the industrial sector, it is considered that digital content delivery is smart enough and suits the nowadays definition of the smart city. Thus, such projects and systems do need two extremities, front end application and a back-end application. Regarding the front-end part, it needs to be adequate for different situations and different user skills. And also, here, we have to be aware that if we want to be practical, software applications and system applications are to be customized [1]. Building a smart environment is a concern, and through this research, we come out with a proven prognostication in a well-designed system, combining the strength of a two-party relationship, machine learning analytics standards, and dedicated and customized hardware to our research. The physical study of AI is not getting the incremental and positive growth it should be, while looking at the digital commerce industry, it is getting strong positive growth in the scientific services. In this research, we have to define our constraints which are used to measure environmental changes in the pre-given parameters; isolated zones, flows of data, etc. Those measures are remarkable points being detected and isolated among different schemas. Thus, the success of such research is a step forward in the context of a secure environment and, more precisely, in the APT detection industry.*

1. Introduction:

The architectural design set by an APT would be similar to the totality of the APT attacks worldwide in terms of numerous criteria. Sever attack done by an APT is referenced to be exposed in terms of several aspects; amongst of them the scope of this research paper. Thus, the APT malware is to be detected similarly at different levels, different network zones, and different behavior scenarios [2, 4].

Besides, we reach to prove through the long research literature review that, a limited list of countermeasure behaviors is a key to limiting an APT behavior within a computerized environment. This limited list of countermeasure behaviors is listed below:

1- Live dataflow capturing.
2- Dataflow analysis in live time.
3- Dataflow behavior classification upon OSI model layers.
4- Dataflow behavior anomalies (classified).
5- Dataflow anomalies continuous detection.

2. Literature Review:

a. Live Dataflow Capturing:

Visualization is widely deployed to explore and analyze complex data. It strives to exploit the capacities of the human perception system, which is very sophisticated, and specifically adapted to locate visual models to interpret a large amount of data. It is not limited to the display of a graph or an image, but it can be seen as a process aimed at graphically representing abstract data, with the aim of identifying trends and correlations which could go unnoticed by looking only at raw or textual data [1, 3, 5, 7].

b. Visualization process

To formalize the visualization process, several models have been proposed. In 1990, Haber and McNabb (Haber and McNabb, 1990) introduced their visualization strategy called “visualization pipeline,” which has three blocks (Figure 2.1):

Filtering: The filtering step prepares the raw input data for processing through the rest of the pipeline steps. It is not only interested in the selection of relevant data but also in operations of data enrichment, interpolation, data cleaning, grouping, and reduction of dimension.

Mapping: This step allows you to map previously prepared data and filter it to visual variables. This is very important because it can influence the efficiency and refinement of graphical representations. If the visual variables are poorly chosen in the sense that the result can be difficult to analyze and vice versa.

Rendering: Rendering allows you to generate the graph from the visual variables from the block.

Mapping: This is where you can use different visualization techniques.

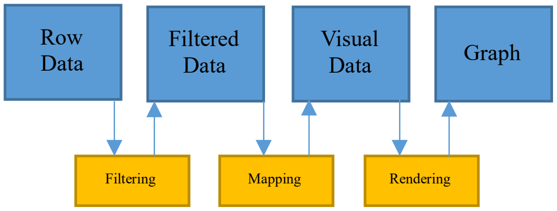


Figure1: Visualization pipeline
Adapted by Aigner et al. (2011)

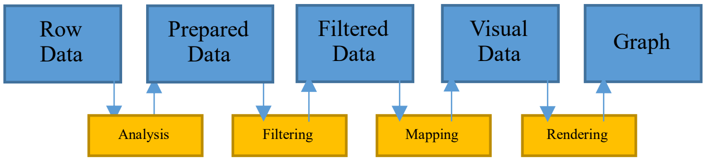


Figure2: Ameliorated visualization pipeline
Adapted by Aigner et al. (2011)

c. Visualization Techniques

Advances made in visualization theory, in computer graphics and its algorithms have given rise to new, more efficient visualization techniques capable of bringing out trends in multivariate data and representing correlations between variables. Despite these advances, and in contrary to the visualization of univariate data, the visualization of multidirectional data encounters several constraints, the most important of which are the three-dimensional representation of the representation space as well as the efficient use of the human visual perception system, which cannot easily design a space of more of three dimensions. For these reasons, the graphical representation of data of four dimensions and more requires the introduction of metaphors, namely colors, shapes, and many others.

The first classification of multidimensional data visualization techniques was proposed by Keim (Keim, 1997, 2002). It distinguishes six categories of visualization techniques:

d. Geometric Techniques

Geometric techniques aim to find interesting transformations of multidimensional data (Keim, 2002). Indeed, they make it possible to project multidimensional data into a new representation space, generally of two dimensions. They are used to process large datasets, mainly to detect outliers and correlations between attributes, notably with the introduction of interaction techniques. A multitude of possibilities of projection in two-dimensional spaces can be imagined, but it is important that the new representations must faithfully reproduce the relevant information contained in the data explored. In addition, techniques from the field of exploratory statistics, typically the scatter matrix, principal component analysis and factor analysis, this category includes other techniques for representing multidimensional data among others the parallel coordinates (Inselberg, 2009).

e. Iconic Techniques

Iconic techniques are based on geometric shapes and icons to represent multidimensional data in a two-dimensional space [4]. They map each observation to a geometric form (glyph) whose visual characteristics (stops, angles, etc.) vary according to the values of the data attributes (Keim, 2002). This approach makes it possible to represent multidimensional data in traditional space. Although the number of dimensions that can be viewed remains limited, these techniques are very useful in this context. When the data attributes are relatively numerous, compared to the dimensions of the representation (two dimensions of the representation space plus the number of visual characteristics of the glyph), the resulting visualization presents visual patterns which vary according to the characteristics of the data. And which can be detected by pre-attentive perception (Keim, 2002). This category includes several techniques, among others, Chernof (Glazar, Marunic, Percic, and Butkovic, 2016), stick figure (Peter J. Sackett, MF Al-Gaylani, Ashutosh Tiwari, and Williams, 2016), and many others.

f. Pixel Oriented Techniques

Pixel-oriented techniques do not allow to visualize only multidimensional data, but also those which are in great quantity. They consist in representing each data value by a colored pixel. For a dataset of dimension n * n, the pixels are used to represent a single observation where the values of each attribute are arranged in a separate window [6]. This class of technique is coming in two main approaches; "Querry-dependent" and "querry-independent technical" (Keim, 1996).

g. Hierarchical Techniques

Hierarchical visualization techniques subdivide the data space into subspaces organized in a hierarchical manner. These techniques are used to represent mainly data that contains a hierarchical structure. This category includes several techniques including TreeMap (Dundas, 2017), Dimensional Stacking (Aigner et al., 2011), and many others.

h. Network Graph Based Techniques

This visualization technique is inspired by the structure of networks in the sense that a graph consists of a set of objects called nodes and which are interconnected by links called "edges" (McGraw_Hill, 2002). This type of visualization makes it possible to bring out the groups (cluster) and makes it possible to discover the trends in the relationships between the different entities. For example, it is used to represent Internet traffic, typically that of social networks.

i. Hybrid Techniques

Hybrid techniques integrates several techniques in one or more windows to produce an expressive graphic representation.

3. METHODOLOGY:

A. Dataflow Analysis in Live Time:

The context of data flow analysis does not stop at a one phase process or approach, yet, it is a multi-phases behavior:

- Classification
- Sampling

b. Traffic Classification

Based on the services line, the classification is to be adopted. Where we can see that the traffic classification goal of the Internet service providers (ISP) is to optimize the use of their network resources and to improve the quality of services offered to customers, or even to meet or exceed their requirements. This can only be achieved with a better understanding of traffic and activity on the networks, especially with the emergence of new applications and SDN (Software Defined Network) [3]. For these reasons, the classification of Internet traffic has aroused particular interest in recent years, from researchers and telecommunications operators. Indeed, traffic classification is a crucial activity in all traffic engineering and network management activities. Typically, quality of service management mechanisms is a direct application of traffic classification. To adequately meet the requirements of the various applications circulating on the network, these mechanisms must refer to the classification of the traffic in order to be able to assign each flow to the appropriate class of service (CoS) and thereby assign it an appropriate priority. The streams belonging to a class are treated differently from the others according to the predefined service quality levels of the similarly, information from the classification and identification of traffic patterns is needed to better design and size networks. With the same level of importance, classification is an essential component in the security system, in particular in the mechanisms of detection of intrusion, unjustified use of network resources or malicious traffic as well as conventional security functions such than firewalls. With technological advances, the emergence of new applications and the popularization of architectures such as SDN (Software Defined Network) and NFV (Network Function Virtualization), bring to light new challenges in network performance and quality of service. They thus motivate the design and development of solutions based on classification and traffic knowledge to achieve this [2].

Two main groups of classification that might act and lead to a better traffic sampling:

1. Taxonomy of traffic classification methods
   a. Classification based on ports
   b. Classification by load inspection
   c. Behavioral approach
   d. Statistical approach
2. Traffic classification and machine learning methods
   a. Decision tree
   b. Decision tree forests (RandomForest Classifier)
   c. Support vector machine (SVM)

c. Traffic sampling

Traffic analysis and network supervision are two essential tasks in the network management process and are an essential activity. This analysis is essentially based on the collection and extraction of the information contained in the packets routed over the network. However, the volume of traffic carried on modern networks is becoming increasingly important; which makes the cost of classification and visualization of traffic more and more important.

A few years ago, the networks offered only relatively low bit rates (around 100MB such as Fiber Distributed Data Interface - FDDI), which could justify the use of probes to collect traffic.

Indeed, these probes copy all the traffic crossing the network node to process it later. This approach is relatively easy to implement and ensures high measurement accuracy, since the measurements are taken from all traffic on a low speed network.

With the evolution of the Internet and the emergence of new technologies that allow the use of broadband networks, new challenges have arisen, including the management of big data. With these changes, traffic analysis using the traditional approach is no longer possible due to the increasing amounts of data routed over broadband networks, which can cause disruptions such as over-saturation of equipment. In fact, collecting all traffic is not only expensive during the processing process, but it can consume a lot of network resources, especially memory resources on the network nodes where traffic is collected. Equipment can be overloaded with this additional task on the one hand, and bandwidth can be significantly affected on the other hand, if the collected data should be sent to a remote backup server [5, 7].

Sampling techniques are around three standardized methodologies seen here:

1. Traffic sampling techniques
   a. Systematic sampling
   b. Random sampling
   c. Adaptive random sampling
2. Standard sFlow
3. NetFlow from Cisco

4. Results:

a) Anomalies detection (APT detection) based on services properties:

APT behavior is an action within an environment, never to diminish such an understanding. Thus, services on the network environment have standard behaviors. Tested services are set and prescribed at the following table:

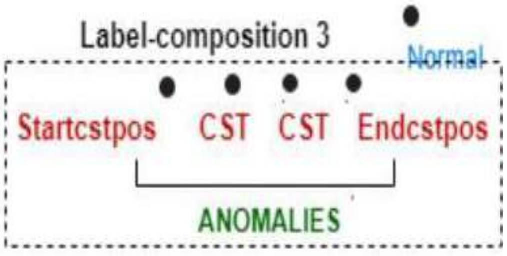
Services		Property							Layer
Name	RFC	synchronous	P2P	CRC	Packet termination	Session	QoS	MTU	
DSL			√	√	Flag	Implicit/Explicit	(1)	X	Physical
ARP	2625		√	X	Flag	Implicit/Explicit	(1)	28-bytes	Link
NAT			√	√	Flag	Implicit/Explicit	(1)		Network
TCP/UDP	793/768	(2)	(2)	√	Flag	Implicit/Explicit	(1)	∞ / 2^{16} - 1	Transport
RPC			√	√	Flag	Implicit/Explicit	(1)		Session
TLS			√	√	Flag	Implicit/Explicit	(1)		Presentation
SMTP	821	X	√	√	Flag	Implicit/Explicit	(1)	∞	Application

(1) Depends on Transport and Physical layers

(2) Those properties are application layer specificities

b) Anomalies detection (APT detection) based on Remarkable Flow Anomalies

Multidimensional visualized flow has presented graphical readings that are clearly presenting an out of range reading. These outliers present an anomaly within the readings. Thus, those anomalies are annotated points within the flow to be utilized for training and testing the automated APT detection solutions. An example of an annotated remarkable point can be seen here:



Network flow remarkable points annotation.

5. Conclusion:

An effective monitoring and management of networks face several obstacles due on the one hand, to the rapid evolution of computer networks which is accompanied by a significant and continuous growth in the quantities of traffic carried, and the other side of the emergence and popularization of many so-called non-standard applications such as P2P applications which cause strong competition in the reservation of network resources for high priority applications such as VoIP which can deteriorate the quality of service offered. Thus, the convergence of traditional telecommunication networks, dedicated to telephony and data transfer networks creates several challenges and new network management issues.

ANNEXE:

RFC: Request for Comment

P2P: Peer to Peer

CRC: Cyclic Redundancy Check up

QoS: Quality of Services

MTU: Maximum Transmission Unit

REFERENCES

[1]. M. de Kunder, “The size of the world wide web,” http://www.worldwidewebsize.com/, accessed: 2014-01-07.

[2]. N. Kshetri, The global cybercrime industry: economic, institutional and strategic perspectives. Springer, 2010.

[3]. F. Valeur and G. Vigna, Intrusion detection and correlation: challenges and solutions. Springer, 2005, vol. 14.

[4]. T. M. technical report, “Targeted attacks and how to defend against them,” http://www.trendmicro.co.uk/media/misc/targeted-attacks-and-how-to-defend-against-them-en.pdf, accessed: 2013-12-20.

[5]. C. Tankard, “Advanced persistent threats and how to monitor and deter them,” Network security, vol. 2011, no. 8, pp. 16–19, 2011.

[6]. P. Wood, M. Nisbet, G. Egan, N. Johnston, K. Haley, B. Krishnappa, T.K. Tran, I. Asrar, O. Cox, S. Hittel et al., “Symantec internet security threat report trends for 2011,” Volume XVII, 2012.

[7]. T. R. Rakes, J. K. Deane, and L. Paul Rees, “It security planning under uncertainty for high-impact events,” Omega, vol. 40, no. 1, pp. 79–88, 2012.