DiNATrAX : a Network Anomalies Detection Framework

Christophe Maudoux, Selma Boumerdassi

DiNATrA$\mathcal{X}$: a Network Anomalies Detection Framework

Christophe Maudoux
CNAM / Cedric Lab
firstname.lastname@cnam.fr
Cnam, 292 rue St. Martin, 75003, Paris, France.

Selma Boumerdassi
CNAM / Cedric Lab
firstname.lastname@cnam.fr
Cnam, 292 rue St. Martin, 75003, Paris, France.

Abstract

Network anomaly detection remains an important research topic in the field of cybersecurity. A network anomaly can be defined as an activity or event that does not correspond to an expected or established traffic behavior. This includes traffic due to cybersecurity attacks (intrusions, malware, DDoS, phishing, etc.), technical failures, or operational errors. Distinguishing between normal and abnormal traffic is essential. Normal traffic is usually defined by a statistical baseline or an expected behavior model, while abnormal traffic deviates from this norm. To detect these anomalies, we propose our framework named DiNATrA$\mathcal{X}$. It is a *generic*, *cyclical*, *adaptable*, and *automatable* methodology based on the use of different *unsupervised* machine learning algorithms. DiNATrA$\mathcal{X}$ is organized into 3 functional blocks. The first block aims to collect and pre-process raw network data. The second block allows for the splitting of the network in order to define logical sectors for analysis. For each of these, a digital signature is generated at regular intervals. These signatures constitute our baseline for comparing network flows. Then, for each of these signatures, we calculate its DNA which allows us to automate the comparison of different signatures. This comparison is performed by the third block which calculates the abnormality distance between 2 consecutive DNAs. If a high abnormality distance is detected, it highlights a variation in network activity and therefore an anomaly which may be correlated with a particular event. Our solution allows us to highligt seven real anomalies correlated to real events from two particular sectors.

Index Terms

Network anomalies detection, digital signature, DNA, Damerau-Levenshtein distance, unsupervised machine learning, activity deviations, network security

I. Introduction

The network anomaly detection framework DiNATrA$\mathcal{X}$ is based on the concept of *DNA* (Digital Network Assessment), which is an evolution of the principle of *digital signatures* detailed in [1]. In this initial approach, network flows are captured and then aggregated in such a way as to constitute well-defined and delimited logical network zones called *sectors*. This step of grouping into sectors is carried out using an unsupervised machine learning algorithm (MLA). Clustering is a method that allows for the identification and grouping of similar data points within larger datasets [2]. Clustering MLAs look for intrinsic common characteristics in the data to extract similarities for grouping purposes. As a result, all data belonging to the same cluster are deemed to be very similar based on the characteristics that comprise them.

Following the results obtained in [1], we chose to produce digital signatures based on 4 clusters. For each of these clusters, a seed was arbitrarily set so that the clustering algorithm always produces the same clusters in the same order when the same data is analyzed. Consequently, it will always be the same data grouped together, thus producing the same cluster with the same identifier (ClusterId) and represented by the same color. Then, in order to detect anomalies related to a given sector, we regularly recalculate the digital signature of the sector at different time intervals. An evolution in the proportions of network characteristics describing traffic leads to a change in the distribution of clusters between two signatures, which reveals a change in activity and therefore potentially a network anomaly.

This approach suffers from two major problems. Since the signatures are calculated at regular intervals but independently, even when fixing the seed, we cannot be sure that the clusters will be computed in the same order between 2 digital signatures. Consequently, similar data between two signatures will be distributed in the same clusters, but the clusters may not have the same identifier. The second problem we faced regarding the analysis of digital signatures is the necessary visual processing that the method requires, and thus the human intervention it involves to determine whether the distribution of clusters is actually different or not. This major drawback effectively prohibits machine processing and, consequently, the automation of the analysis process.

This led to the development of the *DNA* concept implemented by the DiNATrA$\mathcal{X}$ system described in this article, which is structured as follows. First, Section II presents our DiNATrA$\mathcal{X}$ methodology centered around three functional blocks. Section III is a complete implementation of DiNATrA$\mathcal{X}$ carried out within the framework of the CANCAN project [3]. Section IV presents the results obtained and the characteristics that make DiNATrA$\mathcal{X}$ a generic framework.

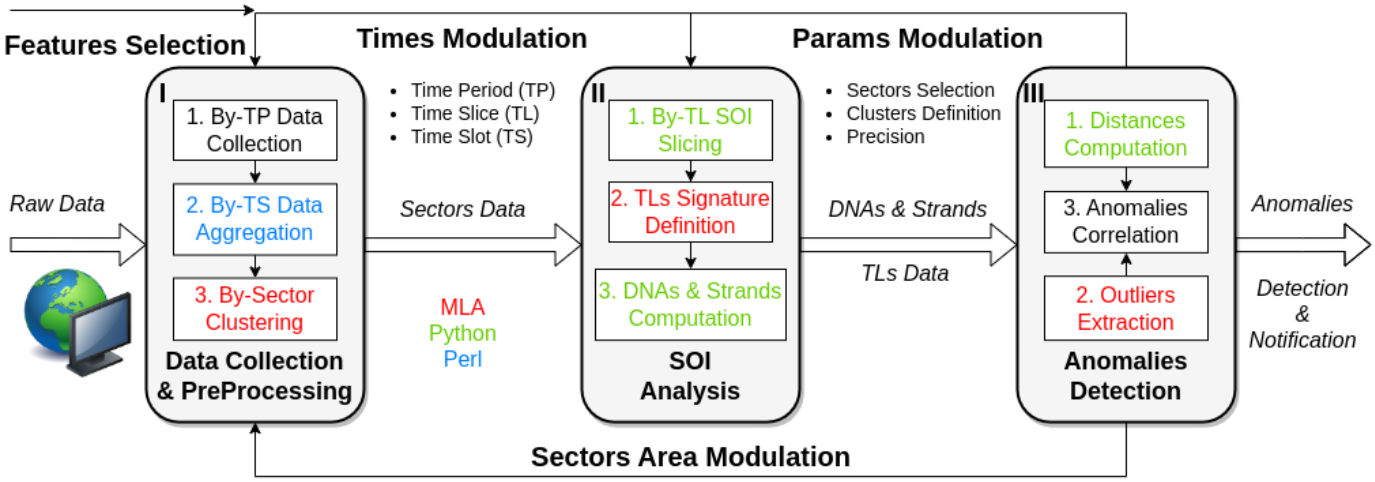


Fig. 1. DiNATrA$\mathcal{X}$ Functional Blocks (FB)

II. DiNATrA$\mathcal{X}$ – Digital Network Assessment & sTRand based Anomalies eXtraction

A. Presentation

The DiNATrA$\mathcal{X}$ network anomaly detection system as described by Fig. 1 is a framework that allows for: *(i)* the collection, preparation & aggregation of network traces *(ii)* their segmentation and analysis by areas of interest, and *(iii)* the generation of alerts if a network anomaly is detected and its possible correlation.

B. Definitions

The DiNATrA$\mathcal{X}$ system defines the notion of *Sector of Interest* (SOI) and employs three different time periods: *Time Period* (TP), *Time Slice* (TL), and *Time Slot* (TS).

1) Sectors & Sectors of Interest (SOI): Detecting anomalies in a LAN-type network can be approached directly and globally, that is, by analyzing all the raw traffic generated by the equipment. However, this approach is not feasible if one wishes to detect anomalies in a larger network, such as a MAN. Indeed, it is not possible to grasp the network as a whole and to directly and fully process all the data circulating in the links. To do this, the network topology must be divided into logical zones or *sectors*. There are no strict or universal rules for defining these logical zones and for this division into sectors. They can be freely delimited based on the topology, the characteristics available during the capture of network frames, hence the generic aspect of DiNATrA$\mathcal{X}$.

Among all these available sectors, only some of them may present a particular interest (such as network periphery, a geographical location, a critical network service, etc.). These sectors of particular interest are called *Sectors of Interest* or SOIs. Depending on the logical division made, its granularity, or the network anomaly to be searched for, it may be interesting to group some of these sectors together. This notion of SOI allows for a level of abstraction above the network topology, enabling the traffic to be described in a *functional* manner.

2) Time Period (TP): This refers to the time period during which network data is collected for analysis. It is the period over which the study will be conducted.

3) Time Slice (TL): The time period TP is divided into time slices TL of equal duration. For each of these TLs, a *digital signature* is generated, and then the *DNA* and its *associated strand* are calculated. This represents the elementary unit of time that allows for the detection of a variation in computer network usage and therefore a potential anomaly.

4) Time Slot (TS): The TS period corresponds to the sampling frequency of network usage. The higher the TS, the greater the granularity and thus the volume of data collected increases, thereby improving the precision with which network activity is described.

C. Functional Blocks

1) FBI — Data Collection & Pre-Processing: This first functional block allows for the collection of network data in raw format using specific tools or equipment. Therefore, it takes raw network frames as input.

Then, it ensures the pre-processing of the collected data for analysis. This pre-processing mainly involves cleaning, consolidating, possibly selecting certain attributes (the 'Features Selection' parameter), or enriching the raw traces (the 'Raw Data' input), and then aggregating them. These aggregated traces are then merged by SOI to allow analysis by well-defined and identified logical zones.

As a result of the FBI, we obtain aggregated and merged data describing the network usage for each of these logical zones or sectors (the 'Sectors Data' output) for the considered TP period.

2) FBII — SOI Analysis: The FBII component represents the heart of the DiNATrA$\mathcal{X}$system. This functional block is responsible for the analysis of aggregated data corresponding to the supervised TP period and related to the selected sector of interest (the 'Sectors Data' input). This sectorial analysis proceeds in four distinct phases: *(i)* the extraction of data from each of the sectors constituting the selected SOI *(ii)* the division of the SOI data into TL time slices *(iii)* the definition of digital signatures for each TL period (TL Digital Signature or TLDS) *(iv)* the calculation for each digital signature of the DNA and its associated strand.

The TLDS digital signatures represent the distribution of data in clusters, carried out by the `K-Means` algorithm, taking into account all the characteristics available in the dataset. From these clusters, we derive what we have named *DNAs (Digital Network Assessments)* and their *associated strands*. Concept of DNA was proposed by [4] and employed by [5].

In the DiNATrA$\mathcal{X}$ framework, DNAs are the string representations of clusters created by `K-Means` based on a particular characteristic denoted as $\mathcal{F}$. This characteristic materializes the network behavior we want to monitor and highlight its evolution, variation over time (data volume exchanged, use by type of protocol, ports involved in communications, returned HTTP codes, TCP flags, etc.). All values or ranges of values that this characteristic $\mathcal{F}$ can take are plotted on the x-axis, and the occurrence of each of these values is represented on the y-axis during the projection of TLDS. The choice of the criterion materializing the occurrence of $\mathcal{F}$ is free.

DNAs consist of as many sequences as there are clusters. A sequence is a string describing the composition of a cluster where each character identifies a value of $\mathcal{F}$ included in the cluster (Seq1 = ADC & Seq2 = BC). The maximum length of a sequence is defined as the *Precision*. The values making up the sequence belong to the set constituted by the different possible values or ranges of $\mathcal{F}$. They are sorted according to their occurrence, which allows obtaining a coded "image" of the clusters. The DNA is then constructed by concatenating the sequences with the '-' character, ordered according to the occurrence of the first value of $\mathcal{F}$ in the sequence. In case of a tie, it is the total occurrences or "Volumes" constituting the sequence that will allow ranking. Indeed, the higher the volume of a sequence, the more representative the cluster described by this sequence will be of $\mathcal{F}$.

To refine the analysis and thus improve the relevance of detection, the *associated strand* of each DNA is calculated. They are deduced from the DNAs and represent the ordered list of occurrences of the different values of $\mathcal{F}$ constituting the DNA. The criterion for measuring occurrences for the values making up the strands is *different* from that used for measuring occurrences during the construction of the DNAs. This allows combining multiple characteristics and *confirming or refuting* the variation.

The following example illustrates how DNAs and their associated strands are constructed. We assume that $\mathcal{F}$ can take the following values: A, B, C, D & E. The number of occurrences for each of these possible values of $\mathcal{F}$ are: A = 10, B = 100, C = 5, D = 50, and E = 1. The sequences are constituted according to the clusters as follows: Seq1 = BDA, Seq2 = BD, Seq3 = AC & Seq4 = DAC. The DNA corresponding to this TLDS will then be: BDA-BD-DAC-AC and its associated strand will thus be: BDAC because 'E' does not appear in the DNA.

Once constituted, all DNAs and strands related to a SOI, as well as the data aggregated by TLDS, are passed to the third functional block in charge of anomaly detection.

Thus, this second block provides to FBIII the list of DNAs and their associated strands for each TL slice for the considered TP period and for the selected SOI (output "DNAs & Strands"). This FB is executed cyclically for each SOI to be analyzed.

3) FBIII — Anomalies Detection: FBIII's role is to detect network anomalies that may have occurred during the TP period and to send an alert if an anomaly is detected. The elements used by this functional component are the DNAs and strands describing the network usage of each SOI (input "DNAs & Strands") as well as the aggregated data by SOI (input "TLs Data"). The use of aggregated data alongside DNAs and strands refines detection and limits the number of false positives. This improves the overall anomaly detection performance of the DiNATrA$\mathcal{X}$system by reducing false positives.

The actual anomaly detection is thus carried out using two different and complementary processes, namely: *(i)* the calculation of the distance between 2 consecutive DNAs and the calculation of the distance between the 2 corresponding strands for all DNAs and strands. From these distances, we calculate the corresponding abnormality distances. *(ii)* the use of an unsupervised MLA which allows us to extract from the TL aggregated data the different aberrant values or *Outliers*. The study of the distribution of the various parameters available in these network data helps to confirm the extracted outliers.

FBIII, if possible, correlates the detected anomalies with events that could explain these changes in behavior within the network and raises and transmits alerts to a supervision tool detailing the aggregated data that enabled the detection of the anomaly. This correlation step is optional as it requires either having implemented a real-time analysis or having a knowledge base of events that may have occurred.

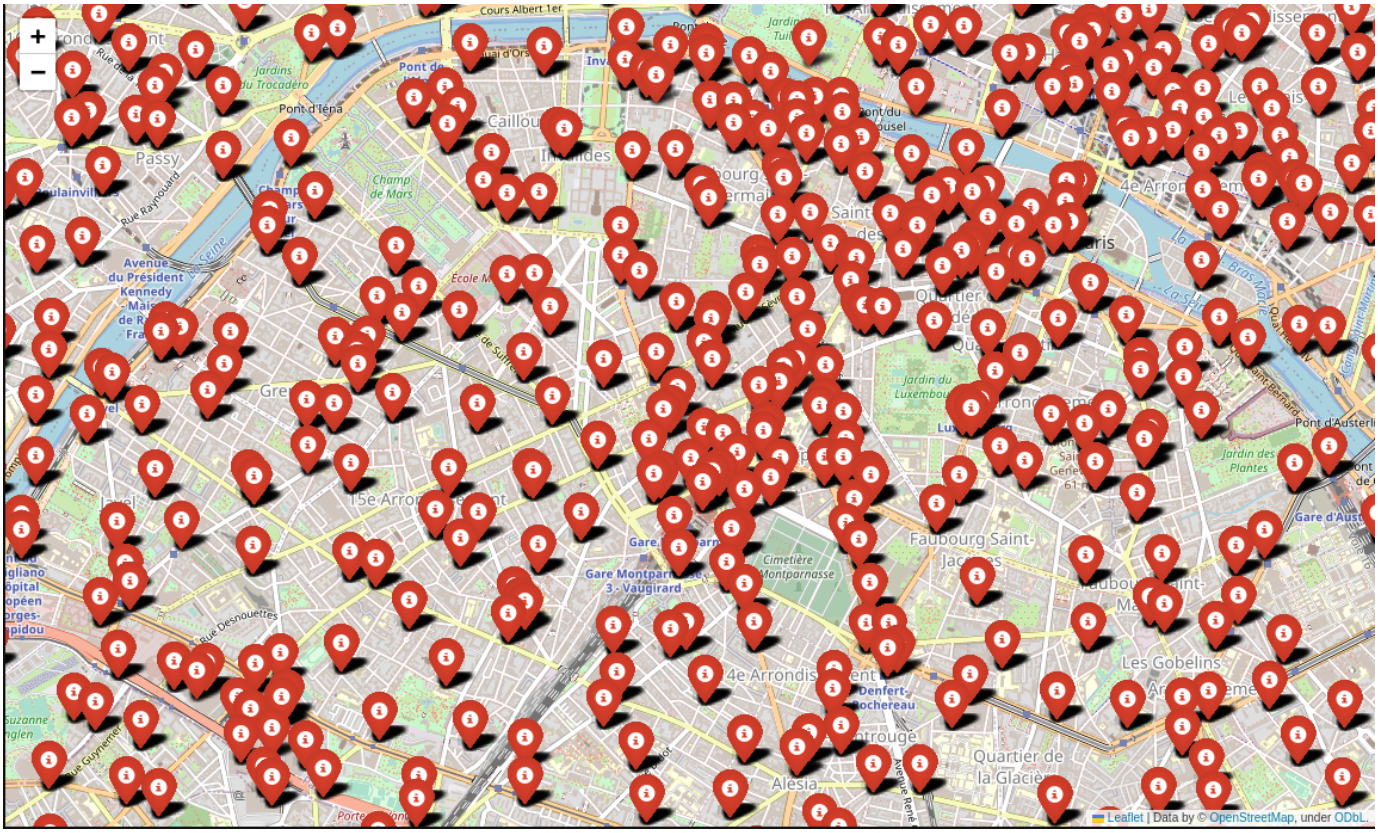


Fig. 2. BTS covering *Center of Paris* Area

TABLE I
'APPGROUP' – APPLICATION TYPE

appGroup	Type	appGroup	Type	appGroup	Type
0 ou A	Unknown	7 ou H	Others	14 ou O	VPN
1 ou B	Web	8 ou I	Control	15 ou P	VVM
2 ou C	P2P	9 ou J	Games	16 ou Q	MMS
3 ou D	Download	10 ou K	Streaming	17 ou R	StreamAVSP
4 ou E	CloudStorage	11 ou L	Chat	18 ou S	Portal
5 ou F	Mail	12 ou M	VoIP		
6 ou G	DB	13 ou N	MailOperator		

III. IMPLEMENTATION

For this study, DiNATrA$\mathcal{X}$ was implemented in `Python` and tested with real data provided by Orange, a major French mobile operator, as part of the CANCAN project [3] and written in the form of a `Jupyter` file downloadable from our repository [6]. These data consist of network flows captured at each Base Transmitting Station (BTS), for which we have GPS coordinates, aggregated by TS over a day and by the 'appGroup' field. This attribute characterizes the type of mobile application used by Orange subscribers. Table I details the correspondence between the references of the application groups and their equivalence in terms of applications. These correspondences were provided by the mobile operator.

These aggregated data thus represent the activity generated by subscribers in the GSM network area covering the entire *Ile-de-France* region. Total number of BTS is equal to 2,736 from which traffic was captured. Fig. 2 is the projection onto a map of some BTS covering the *Paris center* area by using their GPS coordinates included into aggregated data.

A. By-Sector Aggregation

Given the extent of this network and the number of BTS or sources to consider, it is impossible to detect an anomaly by directly analyzing all the data that have been generated. Therefore, this step aims to group BTS into *sectors*. The challenge is to determine which BTS are close enough to be considered *neighboring* in order to include them in the same sector and then merge the data emitted by sector. A sector is thus an area consisting of several BTS close to each other and viewed as a single source.

Given that we have the GPS coordinates of the BTS, the chosen solution for defining these sectors was to use an MLA capable of aggregating geospatial data, namely `DBSCAN`. This method has been tested and implemented by [7], [8] in their work on

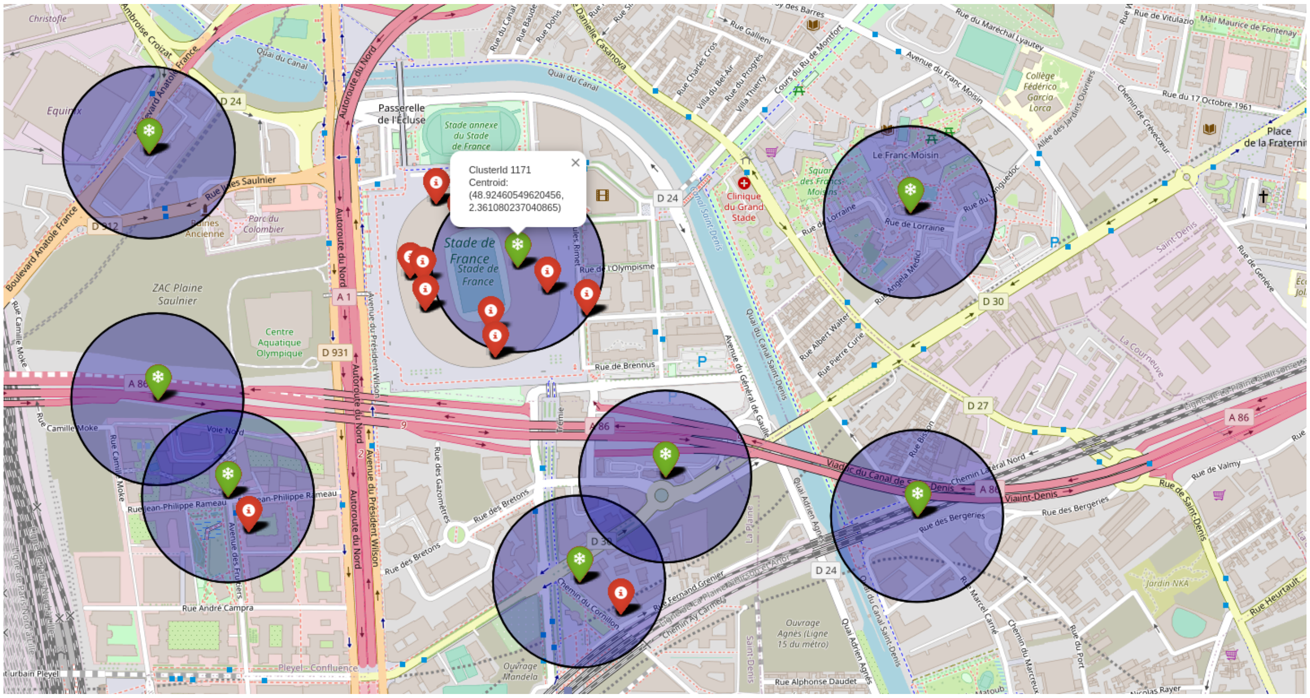


Fig. 3. *Stade de France* SOI & its BTS (red marks)

detecting points of high activity from geographic data from various sources. DBSCAN requires defining several hyperparameters, including the ϵ parameter, which determines the maximum distance at which points can be from each other to be considered as belonging to the same group. It was set at '0.15', which corresponds to a radius of 150 meters. Fig. 3 details the sector of the *Stade de France* (ClusterId = 1171) as well as the BTS that compose it within a 150-meter radius and whose data have been aggregated. An aggregated data sample by TS, 'Cluster,' and 'appGroup' is presented in Table II.

B. By-TL SOI Slicing

A SOI can consist of one or several sectors defined by their 'ClusterId' (see Section II-B1). To analyze the SOIs, it is therefore necessary to extract only the data from the concerned sector(s). This selection of clusters can be done either visually by projecting the sectors onto a map background or by searching for the 'ClusterId' of the sectors to be retained from the GPS coordinates of their *centroid*. Once the SOIs for analysis are selected, the considered analysis period (TP) is divided into equivalent time slices (TL).

In the context of this study, we selected about ten SOIs composed of one or several sectors. Fig. 4 is the projection of the sectors making up the SOI of the *Gare de Lyon*. The aggregated data related to each of these SOIs are then divided by TL to extract their digital signature.

C. TL Digital Signature Extraction

For each SOI, the digital signature of each of the TLs is calculated using the unsupervised MLA K-Means. Machine learning is an important component of data science. Through statistical methods, these algorithms can perform classifications or regressions, which allows for the discovery of essential information in the context of data exploration projects.

According to the works [9], [10], the authors must perform a statistical analysis for each of the parameters contained in the network captures in order to aggregate them. In the context of DiNATrA$\mathcal{X}$, these analyses are carried out by the K-Means algorithm. The data from each TL are grouped into clusters. The distribution of data in the clusters thus formed based on the characteristic $\mathcal{F}$ constitutes the digital signature of each of these TLs: *TL Digital Signature* (TLDS). These signatures

TABLE II
AGGREGATED DATA SAMPLE

TimeSlot	ClusterId	grpDesc	appGroup	Duration	Users	Flows	Packets
2019-05-07	1171	Chat	11	2,230	4365	13,480	10,331
2019-05-09	1171	Chat	11	2,342	4324	15406	11,660
2019-05-08	1171	Web	1	595.825	2	4	765
2019-05-01	1171	Web	1	1,546.769	3	3	971
2019-04-29	1171	VVM	15	18.084	1	1	145

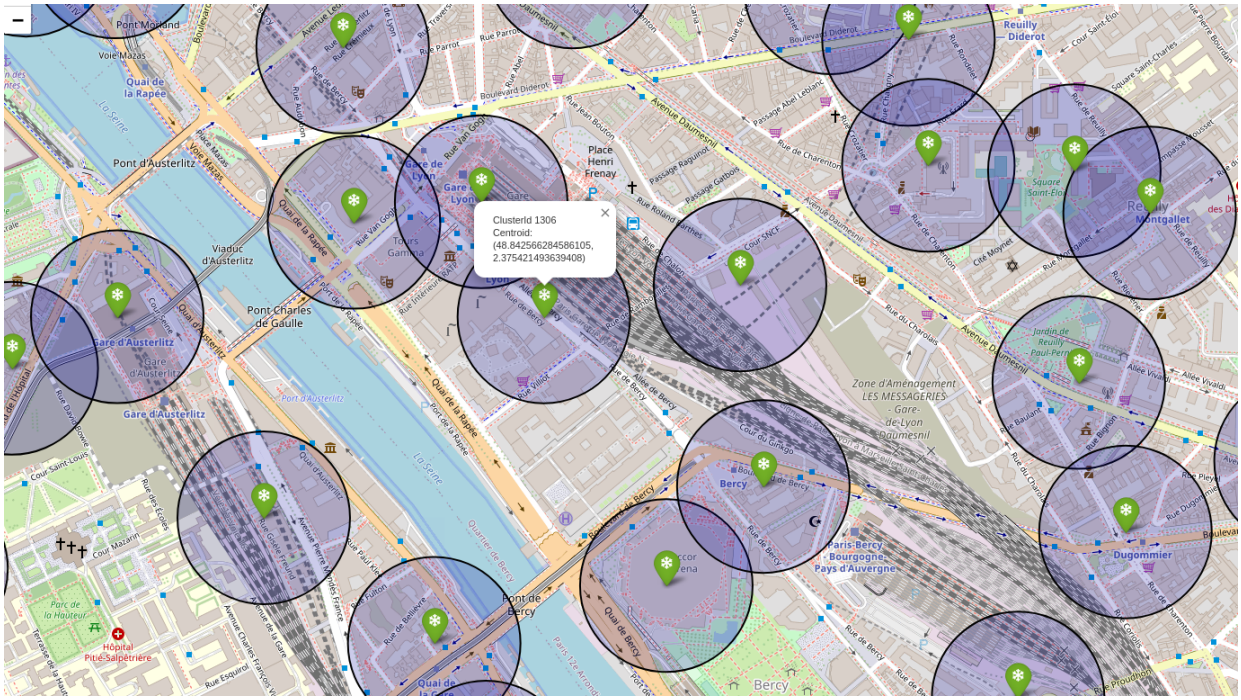
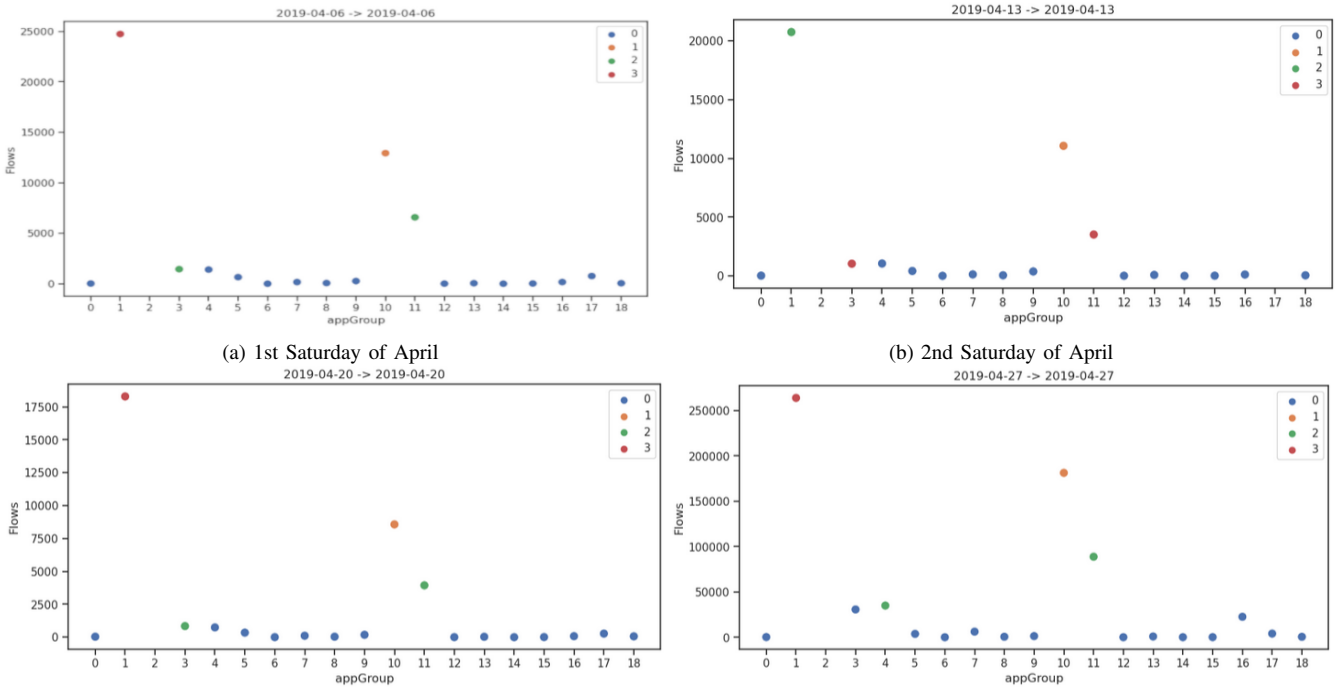


Fig. 4. *Gare de Lyon* SOI composed of 3 sectors



(a) 1st Saturday of April

(b) 2nd Saturday of April

(c) 3rd Saturday of April

(d) 4th Saturday of April

Fig. 5. *Stade de France* SOI TLDS (TP=1 month & TL=24H)

can be materialized by projecting the clusters onto a plane whose x-axis represents the characteristic $\mathcal{F}$ to be analyzed and the y-axis a parameter characterizing its evolution. For the analysis of CANCAN data, we chose 'Flows' on the x-axis and $\mathcal{F} = appGroup$ to detect anomalies related to changes in user behavior materialized by the type of application used. Fig. 5 represents the digital signature of the 4 Saturdays of April 2019 for the SOI of the *Stade de France*."

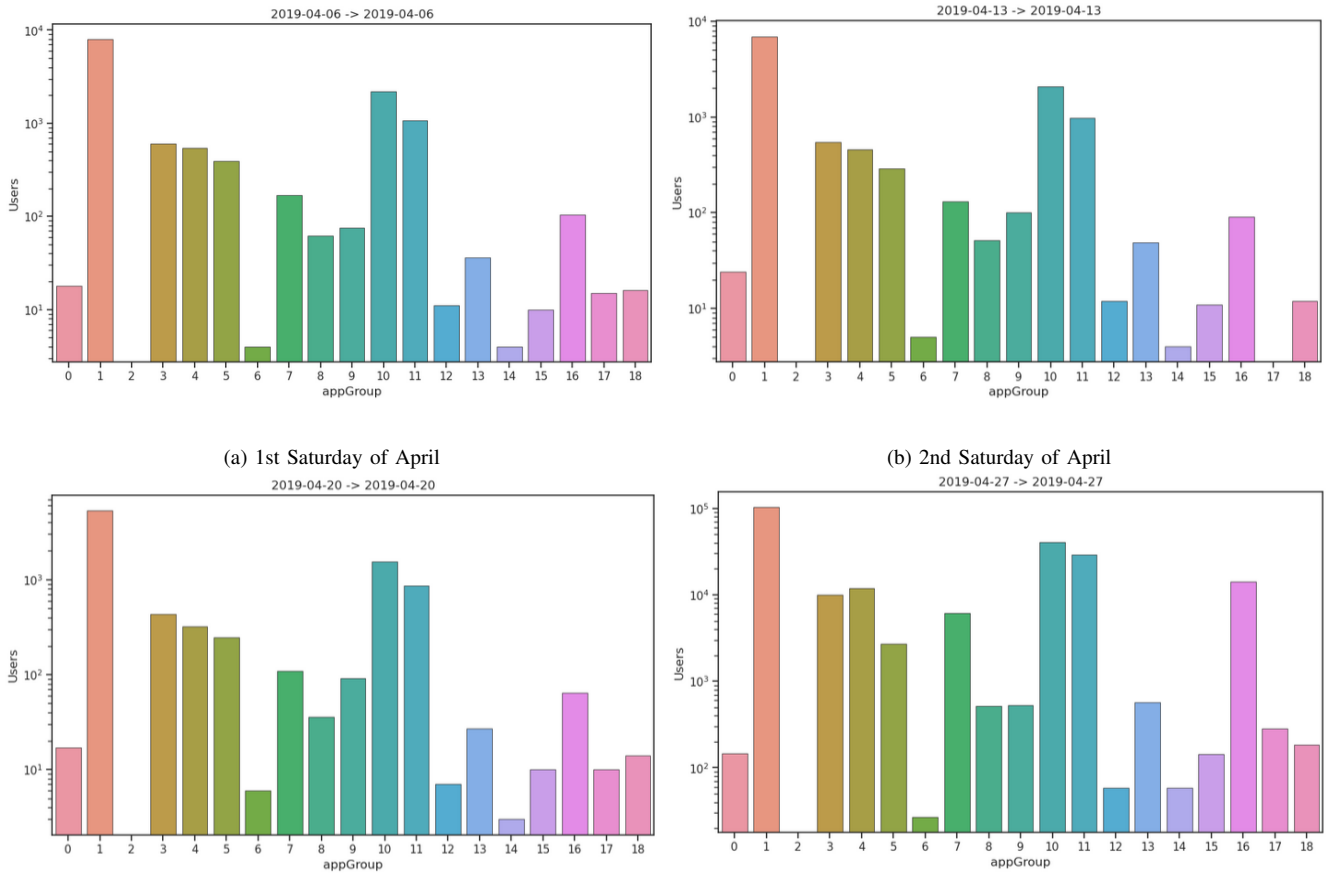
Once all the TLDS are generated, the next step is to compare them two by two. If an evolution in the distribution of clusters

is observed between two consecutive TLDS, this implies that there has been a variation in the characteristic $\mathcal{F}$ ('appGroup'). This comparison requires human intervention as it requires visual interpretation. Indeed, as discussed in Section I, the numbering of clusters and therefore the color associated with them may vary from one signature to another. As a result, although the distribution of data is identical between two signatures, they could be interpreted as different if an automated comparison was employed.

Hence the concepts of 'Digital Network Assessment' (DNA) and 'Strand' defined and used by DiNATrA$\mathcal{X}$.

D. DNAs & Strands Computation

For each signature, what we have called the *DNA* and its *associated strand* are calculated. The calculation of the DNAs provides us with a first indication of the existence or absence of a potential anomaly. Then, the *abnormality distance* of each strand is calculated. The letters making up the strands are sorted according to a different parameter representing activity, here the number of users ('Users') per 'appGroup'. These two pieces of information allow us to confirm or deny the presence of an anomaly in the considered SOI when moving from TL_n to TL_{n+1}. This calculation is performed for each TLDS of TP of the considered SOI and repeated for each of the SOIs to be studied. Results obtained for the *Stade de France* SOI are presented and explained below based on Table III and Fig. 6.



(a) 1st Saturday of April

(b) 2nd Saturday of April

(c) 3rd Saturday of April

(d) 4th Saturday of April

Fig. 6. *Stade de France* SOI Activity

TABLE III
Stade de France SOI DNAs & STRANDS

n	Begin	End	DNA_n	$Strand_n$	Volumes
0	2019-04-06	2019-04-06	B-K-LD-EFHQJI	BKLDEF	59.93-16.49-12.57-10.15
1	2019-04-13	2019-04-13	B-K-LD-EFHJQI	BKLDEF	58.98-17.58-12.95-9.49
2	2019-04-20	2019-04-20	B-K-LD-EFHJQI	BKLDEF	58.33-16.93-14.19-9.52
3	2019-04-27	2019-04-27	B-LE-K-QDHFNJ	BKLQED	46.66-18.72-18.43-15.54

TABLE IV
Stade de France SOI ANORMALITY DISTANCES

n	Begin	End	DNA	Strand	DNAD	STAD	ANOD
0	2019-04-06	2019-04-07	BK-B-LD-EFHQIJ	BKLDEF	7	2	9
1	2019-04-13	2019-04-14	B-K-LD-EFHQJI	BKLDEF	3	0	3
2	2019-04-20	2019-04-21	B-K-LD-EFHJQI	BKLDEF	1	0	1
3	2019-04-27	2019-04-28	BK-B-LED-QEDHFI	BKLQED	7	2	9
4	2019-05-04	2019-05-05	BK-B-LDE-DEFHQJ	BKLDEF	5	2	7
5	2019-05-12	2019-05-13	BK-B-LQEDFH-LED	BKLQED	7	2	9

For the analysis of the SOI *Stade de France*, we divided the TP period, corresponding to April 2019, into 24-hour slices to detect anomalies related to a change in the daily activity of users registered in this area of Orange's mobile network. We chose to study the four TLs Saturdays contained in TP period.

TL_0 contains data related to the network activity of Saturday, April 6, 2019. Fig. 5a represents $TLDS_0$, the digital signature of TL_0. DNA_0, the DNA corresponding to this signature, consists of four sequences because *K-Mean* is configured to create digital signatures by distributing data into 4 clusters.

Seq_3 is the expression of $cluster_3$ according to the characteristic $\mathcal{F}$ here 'appGroup'. $Seq_3 = B$ because $cluster_3$ consists only of 'appGroup' 1 which corresponds to the letter 'B' representing 'Web' applications as detailed in Table I. $Seq_1 = K$ because $cluster_3$ consists only of 'appGroup' 10. $Seq_2 = LD$ because $Cluster_2$ consists of 'appGroup' 3 & 11. $Seq_0 = EFHQJI$ because $Cluster_0$ consists of 'appGroup' 0, 4 to 9 & 12 to 18. Sequences are truncated to a length equal to the $Precision$. The greater the $Precision$, the greater the number of different values of $\mathcal{F}$ considered. This means that application groups with increasingly lower activity will be considered. The letters forming the sequences ('appGroupL') are sorted according to the characteristic describing the activity of the 'appGroup', here 'Users' which represents the number of users per type of application, activities detailed by Fig. 6a.

Finally, for each sequence, the volume of data represented by the sequence is calculated, expressed as a percentage (59.93-16.49-12.57-9.68). Sequences are then sorted in descending order of volumes to construct the DNA hence: $DNA_0 = Seq_3 - Seq_1 - Seq_2 - Seq_0 = B - K - LD - EFHQJI$. The strand $Strand_0$ associated with the DNA DNA_0 is made up of all the values of $\mathcal{F}$ contained in DNA_0, sorted in descending order of activity and truncated to the length $Precision$: $Strand_0 = BKLDEF$.

This process is repeated for each of the TLs. The results obtained are detailed in Table III. We can deduce that: *(i)* the strands $Strand_0$ to $Strand_2$ are identical *(ii)* the DNA_0 to DNA_2 are almost identical except for the application groups 'J' and 'Q' representing 'Games' and 'MMS' swapped in DNA_0 *(iii)* the activity volumes by sequence are of the same magnitude for periods TL_0 to TL_2 *(iv)* the DNA, its strand, and the activity volumes of period TL_3 are different, which highlights a network anomaly for this period compared to periods TL_0 to TL_2.

According to Table I, we can conclude that on Saturdays, April 6, 13, and 20, 2019, at the SOI of the *Stade de France*, users primarily used their mobile phones for 'BKLDEF' activities: Web (60%), Streaming (17%), Chat & Download (13%), CloudStorage, Mail, and others (10%). For Saturday, April 27, 2019, the activity corresponded to 'BKLQED': Web (47%), Chat & CloudStorage (19%), Streaming (18%), MMS, Download, and others (15.5%).

E. Abnormality Distances Computation

Now, it is necessary to confirm detected anomalies. To limit false positives and unambiguously identify relevant anomalies, we determine for each DNA and each strand what we call the *abnormality distances*. These distances allow us to quantify the detected anomalies and to send a notification only for certain ones based on predefined thresholds. These abnormality distances are based on the principle of DAMERAU–LEVENSHTEIN distance, denoted DL($String_1, String_2$).

For each DNA_n, the following abnormality distances are calculated: $DNAD_n = DL(DNA_{n-1}, DNA_n)$ and $STAD_n = DL(Strand_{n-1}, Strand_n)$ where 'n' represents the index of the concerned TL period. From these distances, the total abnormality distance is deduced: $ANOD_n = DNAD_n + STAD_n$. This distance $ANOD_n$ materializes the *amount of abnormality* of the time slice TL_n compared to the previous time slice TL_{n-1} and therefore formalizes the rate of variation in network activity by type of applications of users during the transitions from TL_{n-1} to TL_n during the TP period for the considered SOI.

From Table IV, we can draw the following observations: 1) the slices TL_0, TL_3, TL_4 & TL_5 show a high abnormality distance 2) the slices TL_0 to TL_2 & TL_4 have identical strands 'BKLDEF' 3) the slices TL_3 & TL_5 have identical strands: 'BKLQED' 4) the slices TL_0 to TL_2 have very similar DNAs 5) DNA_4 is a mix between DNA_2 & DNA_3 6) The abnormality distances of TL_0 & TL_5 are identical or close 7) the slices TL_3 & TL_5 have almost identical DNAs. The difference is due to the application group 'L' (Chat) appearing in the sequence Seq_3 of DNA_5

From observations '2' and '4', we can deduce that the periods TL_0, TL_1, TL_2, and TL_4 present identical network activity. From observations '1' to '3', we can deduce that the periods TL_3 and TL_5 present a major evolution in user activity, which materializes two anomalies. From observations '2' and '6', we can deduce that the transition between periods TL_5 and TL_0 is a return to the reputed "normal" or reference state of network activity. From observations '1' and '5', we can deduce that

TABLE V
DETECTED ANOMALIES

AnoId	SOI	TimeSlot	ClusterId	ANOD	Event	Ref
1	*Montmartre Cmy*	2019-04-08	962	7	Funeral	
2	*Montmartre Cmy*	2019-05-02	962	6	D. Rivers' Funeral	[12]
3	*Notre Dame de Paris*	2019-04-15	1037+1060	4	Fire	[13]
4	*Notre Dame de Paris*	2019-04-16	1037+1060	4	Fire	[13]
5	*Stade de France*	2019-04-27	1168	8	French Cup Final	[14]
6	*Stade de France*	2019-05-12	1168	6	Metallica Concert	[14]
7	*Gare de Lyon*	2019-04-15	All	7 / 6	Orange 4G Outage	[15]
Detected Anomalies (Cmy = Cemetery)						

the period TL_4 presents an anomaly due to a return to the activity normally observed for periods TL_0, TL_1, TL_2. Finally, according to observation '7', the network activity during these two TLs was identical. This means that mobile users used almost the same applications and in the same proportions during these periods.

IV. CONCLUSION & FURTHER WORKS

From the analysis of these few SOIs, it appears that DiNATrA$\mathcal{X}$ has enabled us to identify seven anomalies. These can sometimes be linked to specific events that explain them. To do this, we extract the outliers of each SOI from the aggregated data using the unsupervised MLA `Local Outlier Factor`. LOF is an algorithm commonly used to detect unusual network activities [11]. To confirm these outliers, a study of the distribution of the aggregated data is carried out. This study focuses on the data of the concerned TLs and on all the data of the SOI. Table V summarizes the anomalies linked to the analyzed SOIs and the probable events that caused them. From this table, we can affirm that we have been able to detect six anomalies related to a change in user activity. The seventh anomaly is more likely related to a network incident as it is present in two different SOIs.

This organization of DiNATrA$\mathcal{X}$ into three distinct and independent functional blocks ensures us a framework that is transposable regardless of the network topology to be analyzed and agnostic to the type of anomalies sought: *(i)* the network characteristics extracted and analyzed are freely chosen and depend only on the context in which it is deployed *(ii)* the time periods, the parameter values can be adapted to best meet the needs *(iii)* the granularity of the sectors can be modulated to adapt to the concerned network. We therefore plan to test DiNATrA$\mathcal{X}$ with other datasets and to deploy it in "real life".

REFERENCES

[1] C. Maudoux and S. Boumerdassi, "Network Anomalies Detection by Unsupervised Activity Deviations Extraction," in *2022 Global Information Infrastructure and Networking Symposium (GIIS)*, Sep. 2022, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9937022

[2] A. Jung and I. Baranov, "Basic Principles of Clustering Methods," Dec. 2019. [Online]. Available: http://arxiv.org/abs/1911.07891

[3] "Content and Context based Adaptation in Mobile Networks." [Online]. Available: https://anr.fr/Project-ANR-18-CE25-0011

[4] J. Blakes, O. Raz, U. Feige, J. Bacardit, P. Widera, T. Ben-Yehezkel, E. Shapiro, and N. Krasnogor, "Heuristic for Maximizing DNA Reuse in Synthetic DNA Library Assembly," *ACS Synthetic Biology*, vol. 3, no. 8, pp. 529–542, Aug. 2014. [Online]. Available: https://doi.org/10.1021/sb400161v

[5] P. Siyari, B. Dilkina, and C. Dovrolis, "Evolution of Hierarchical Structure & Reuse in iGEM Synthetic DNA Sequences," Jun. 2019. [Online]. Available: http://arxiv.org/abs/1906.02446

[6] C. Maudoux and S. Boumerdassi, "DiNATraX_CANCAN." [Online]. Available: https://kaggle.com/code/christophemaudoux/dinatrax-cancan

[7] U. Angkhawey and V. Muangsin, "Detecting Points of Interest in a City from Taxi GPS with Adaptive DBSCAN," in *2018 Seventh ICT International Student Project Conference (ICT-ISPC)*, Jul. 2018, pp. 1–6.

[8] G. Boeing, "Clustering to Reduce Spatial Data Set Size," Mar. 2018. [Online]. Available: https://osf.io/preprints/socarxiv/nzhdc/

[9] E. H. M. Pena, M. V. O. de Assis, and M. L. Proença, "Anomaly Detection Using Forecasting Methods ARIMA and HWDS," in *2013 32nd International Conference of the Chilean Computer Science Society (SCCC)*, Nov. 2013, pp. 63–66.

[10] E. H. M. Pena, S. Barbon, J. J. P. C. Rodrigues, and M. L. Proença, "Anomaly detection using digital signature of network segment with adaptive ARIMA model and Paraconsistent Logic," in *2014 IEEE Symposium on Computers and Communications (ISCC)*, Jun. 2014.

[11] J. Auskalnis, N. Paulauskas, and A. Baskys, "Application of Local Outlier Factor Algorithm to Detect Anomalies in Computer Network," *Elektronika ir Elektrotechnika*, vol. 24, Jun. 2018.

[12] "Obsèques de Dick Rivers," May 2019. [Online]. Available: https://www.leparisien.fr/culture-loisirs/musique/obseques-de-dick-rivers-c-etait-le-plus-rockeur-des-rockeurs.php

[13] "Retour sur l'incendie de Notre-Dame de Paris." [Online]. Available: https://revivre-notre-dame.fr/incendie-cathedrale-notre-dame/incendie-notre-dame-de-paris/

[14] "Evènements passés." [Online]. Available: https://www.stadefrance.com/fr/billetterie/archives

[15] M. Turcan, "Panne chez Orange : des problèmes de connexion Internet et 4G sur toute la France," Apr. 2019. [Online]. Available: https://www.numerama.com/tech/panne-4g-orange.html