



HAL
open science

Unsupervised Anomaly Knowledge Flow: a Digital Signatures Extraction Approach

Christophe Maudoux, Selma Boumerdassi

► **To cite this version:**

Christophe Maudoux, Selma Boumerdassi. Unsupervised Anomaly Knowledge Flow: a Digital Signatures Extraction Approach. 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM), Oct 2023, Istanbul, Turkey. pp.1-6, 10.1109/WINCOM59760.2023.10323022 . hal-04316873

HAL Id: hal-04316873

<https://hal.science/hal-04316873v1>

Submitted on 30 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License


Unsupervised Anomaly Knowledge Flow: a Digital Signatures Extraction Approach

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—Various machine learning or clustering techniques are applicable for identifying anomalous activities or particular events in networks by analysing data flows. In this paper we present our networks anomalies detection system which is based on a framework that we named *Unsupervised Anomaly Knowledge Flow*. Our approach consists of aggregating pre-processed network flows into well-known areas named sectors. For each sector, data describing users activity are aggregated and split into different equal time-periods. After this step, an unsupervised clustering algorithm is employed to extract the *digital signatures* defining sectors activity. If a specific sector signature for one specific period differs from others, it means that a network anomaly relative to users activity has been detected. A last step is performed to associate highlighted anomalies with their respective events. This framework originality are its *generic*, *cyclic* and *fractal* aspects. Our experiments have been conducted by using a *real dataset* captured and provided in 2019 by a major French mobile operator. Our proposed knowledge flow is able to detect anomalies related to real crowded events like the *Notre-Dame de Paris* fire, concerts or soccer matches. For this study, sectors have been computed by using geographic coordinates defining the base transmitting stations, and anomalies are reliant on network activity features.

Index Terms—Geohash, network anomalies detection, activity deviations, digital signature extraction, unsupervised machine learning, crowded events

I. INTRODUCTION

Detecting suspicious or abnormal traffic in networks is still a critical need for security purposes or resources allocation [1]. *Network anomalies* are traffic activity that differ from what is currently observed. Point out network anomalies means that you have to know what is an expected behaviour. Related studies have been conducted by monitoring network traffic and analysing features or by parsing their payload like [2], [3].

Those approaches suffer from prohibitive downsides: (i) intrusion detection process is based on rules built over basic network features or attributes like HTTP methods or source and destination ports. Furthermore the rules might be updated according to network evolutions [4] (ii) each solution is deployed into specific locations (iii) each location is isolated and not linked so collected data has to be aggregated and correlated to highlight flows relating to a particular anomaly.

The first weak point could be minimized by subscribing to an update or intelligence service. But you still depend on the update scheduling and provider. Secondly locations

selection is not exhaustive and could be difficult to perform efficiently due to network architecture or complexity. The last and most serious drawback is the lack of a global overview. Some anomalies are devious. The only way to detect them is to aggregate flows coming from different IDS. This can be done by deploying a well-managed Security Incident and Event Management (SIEM) platform which is a collection of cybersecurity or analysis components used to monitor network traffic or resources by collecting low level Key Performance Indicators (KPIs), to provide reports and to trigger alarms [5].

In this paper, we propose a framework for detecting behaviour changes related to mobile users activity. For this purpose, we compute *digital signatures* (DiSi) that are like pictures taken at a specific moment in time and for a particular network slice or zone. This article which is an extension of our previous works [6] describes our network anomalies detection method named "Unsupervised Anomalies Knowledge Flow" using Machine Learning Algorithms (MLAs) to extract *DiSi* and highlight *outliers*. Our framework offers three main strong aspects that leads to a versatile model. Concept of digital signatures can be: (i) applied to different networks by example LAN, WAN; *generic* (ii) scheduled at different moments in time depending on supervised or monitored phenomenon: *cyclic* (iii) computed for different network areas or zones: *fractal*.

After introducing overall context in Section I, we describe our proposed framework in Section II. Section III presents the dataset employed for defining our methodology. Section IV exposes our implementation and summarizes obtained results. Section V concludes by suggesting possible improvements and we define our planned further work.

II. UNSUPERVISED ANOMALIES KNOWLEDGE FLOW

Our study focuses on mobile traffic captured from Orange relays, a french mobile operator, covering the *Île-de-France* region. For this zone, total amount of data to analyze represents over 190 GB that can not be processed and analyzed directly. So, we defined an unsupervised anomaly detection methodology called "Unsupervised Anomalies Knowledge Flow" which is broken down into four distinct phases: (i) raw data aggregation using parsers (ii) digital signatures extraction from aggregated streams and (iii) outliers detection both using an unsupervised MLA (iv) anomalies correlation.

The necessary initial pre-processing phase aims to analyze, verify, clean and aggregate network flows. Goal here is to reduce the total amount of data to be analyzed, without any loss of information, so that the underlying user behaviors can be extracted. This aggregation phase is performed in 2 steps:

(i) aggregate by Base Transceiver Station (BTS) the raw network flows captured at the level of the various antennas making up these BTS: the *By-Site Aggregation* step (ii) merged the data corresponding to the network traffic generated and captured at each site to define what we have called *sectors*: the *By-Sector Aggregation* step.

This first By-Site and then By-Sector aggregation phase, while preserving the network characteristics of the flows exchanged, will then enable us to extract what we call *digital signatures* from these detailed aggregated streams. In next phases, they will be used for highlighting network anomalies.

A. Data Pre-Processing

1) *By-Site Aggregation*: This first aggregation step consists in sorting, cleaning, transposing and merging the raw network flows captured at each of the BTS deployed by Orange in *Île-de-France*. These flows, captured as part of the CANSAN project, are made available in the form of 25 '.csv' files. The data contained in these files was captured from each of the 4G antennas installed at each BTS. In fact, a BTS may comprise several radio antennas. We therefore start by merging the flows from antennas belonging to the same base station to define what we call *sites* and thus obtain aggregated flows representing the site's network activity. The type of network activity generated by users depends on the mobile applications used. This information is provided by the 'PortApp' field. For our use case, the application used is too detailed a piece of information. We have therefore chosen to base our study on the application group or 'AppGroup'. For example, we simply need to know whether the user is browsing the Internet (web) or sending electronic messages (e-mail), but not which browser or e-mail client is being used. This step of merging the raw data for each site is represented by red triangles in Fig. 2.

2) *By-Sector Aggregation*: This second step consists of aggregating the merged data from sites identified as *neighbors* to form *sectors*, as represented by the blue rectangle in Fig. 2. Aggregation by sector aims to select specific sectors or areas for more detailed analysis. This step of aggregating the merged data from each site to form sectors presents three difficulties: 1) How to define sectors (size, shape, on what basis)? 2) How to know if sites are neighbors? 3) How to determine whether the site should be included in the sector?

To address these 3 difficulties, we opted for the GEOHASH encoding system [7]. It is a public domain geocoding system invented in 2008 by GUSTAVO NIEMEYER that uses GPS coordinates to encode a geographic location into a short string of letters and digits. It is a hierarchical spatial data structure that subdivides the Earth's surface into a hierarchical grid. Geohashes offer properties such as arbitrary precision and the ability to progressively eliminate characters from the end of the code to reduce its size, gradually losing precision. As a

consequence of the gradual degradation of Precision, nearby locations often (but not always) have similar prefixes: the more similar a shared prefix is, the closer the two locations are.

To determine the geohash of a point, you need to have its latitude and longitude coordinates into GPS system. However, the coordinates available in the CANSAN dataset are defined in the LAMBERTII system. So, in order to calculate the geohash associated with each site, we first had to convert these coordinates into the GPS system using an online converter.

With the GEOHASH encoding system, the area defining a sector depends on its associated geohash length as depicted by Fig. 1. This length is also known as *Precision*. The higher the Precision, the smaller the corresponding sectorial zone, as shown in Figs. 1a to 1c. To determine the most appropriate Precision value, different lengths from 5 to 7-character geohashes were tested. A 6-character geohash was chosen because its Precision corresponds to well-defined and comprehensible study locations such as *Notre-Dame de Paris* or *Stade de France*. A geohash with *Precision* = 6 defines a cell with a height of 1.2 km and a width of around 0.6 km, as shown in Fig. 1b and detailed by Table I.



(a) P=5 (u09wm) (b) P=6 (u09wmd) (c) P=7 (u09wmdw)

Fig. 1. Sector size depending on Precision

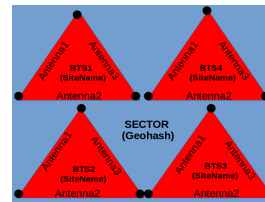


Fig. 2. By-Site (red triangles) & By-Sector (blue rectangles)

TABLE I
PRECISION &
CORRESPONDING DIMENSIONS

Precision	Length	Width
4	39	19.5
5	4.9	4.9
6	1.2	0.6
7	0.153	0.153
8	0.038	0.019

B. Digital Signatures Extraction

The DiSi extraction phase consists, for each sector, in grouping the streams aggregated in the previous step into "clusters". This grouping of aggregated data into clusters is repeated periodically, for time intervals of equivalent duration. Data is gathered on the basis of attributes calculated during the aggregation stage. The way in which the data is distributed within the various clusters per period is what we have termed the *digital signature* of a period or time slice. If, for the same sector, the signature of a period differs from the others, it is considered suspicious and highlights a behavior of users on the network considered abnormal because deviating from the others. This step is carried out using the unsupervised machine learning algorithm X-Means.

The X-Means algorithm is a variant K-Means that makes it possible to refine the assignment of groups by iteratively trying to subdivide them, keeping only the best results [8]. Data are split into K mutually exclusive clusters in such a way that the data in each cluster are as close as possible to each other, but as far apart as possible from the data forming the other clusters. Each cluster is characterized by its central point named *centroid*. It acts of the point which coordinates are obtained calculating the average of all coordinates points sampling attributed to each cluster.

C. Outliers detection

This third phase aims to detect network anomalies by highlighting outliers among the detailed aggregated data set. It is performed using the Local Outlier Factors algorithm [9]. LOF is based on the concept of local density, where locality is given by the k nearest neighbors. By comparing the local density of an object with the local densities of its neighbors, it is possible to identify regions of similar density and points whose density is significantly lower than that of their neighbors. The latter are then considered outliers.

D. Anomalies correlation

Once the outliers have been identified, the final phase of our methodology consists in correlating the extracted anomalies with real facts or events that could justify these changes in network activity, such as outages or crowd movements. This step enables us to confirm and explain the anomalies.

III. EMPLOYED DATASET

The dataset employed for this study has been collected and provided by Orange as part of the CANCAN project [10]. This project aims to bring about the next generation of mobile networks with cognitive capabilities. In fact, mobile traffic is characterized by a strong contextual and content heterogeneity.

Gathered data characterize the mobile traffic and application usage over one year in France. They are composed of different fields: **PortApp** defines the mobile application type used by clients, **LocInfo** is an identifier defining the antenna receiving data, **Coord_X & Coord_Y** correspond to site coordinates in LAMBERTII system, **SiteName** is the site's name, **TimeSlot** is date and time in 30mn slots when traffic has been captured, **Duration** corresponds to TCP session duration, **nPktUp & nPktDn** are total number of packets received and sent during TCP session, **Flows** details number of traffic flows exchanged during TCP session, **Users** is the number of distinct users.

IV. EXPERIMENTS & RESULTS

For the purposes of this study, we have chosen to analyze a period corresponding to three months worth of data. To do this, we extracted from the dataset the raw traces generated between 16/03/2019 and 14/06/2019 inclusive. These network flows were captured at regular 30-minute intervals by the mobile operator using network probes placed at each antenna. A snippet describing a few lines of raw data supplied as part of the CANCAN project is provided by Table II. Some details have been removed for confidentiality reasons.

TABLE II
SNIPPET OF CANCAN RAW DATA

PApp	LocInf	CoordX/Y	SiteName	TimeSlot	PktU/D	Dur	Usr	Flw
65734	cdb04	143 / 023	U_PEL	19-05-14 21:00	49/180	58	1	1
66333	cdb02	143 / 023	U_PEL	19-05-14 11:30	4/83	2	1	1
65759	cdb05	143 / 023	U_PEL	19-05-13 08:30	10/468	633	1	2
66327	7a280	172 / 971	U_TRIB	19-04-11 10:30	11/849	5	1	3
65701	1cc5d	473 / 054	UNION	19-04-11 16:30	1/97	24	1	1
65745	1cc5d	473 / 054	UNION	19-04-11 07:30	8/338	23	1	2
65745	1cc5d	473 / 054	UNION	19-04-11 12:30	13/338	51	1	3
66358	1cc5d	473 / 054	UNION	19-04-11 09:30	5/419	27	1	1

Data extracted from 'df_U.csv' file (some values have been truncated for commercial privacy)

As specified in Section III, the raw data consists of 11 fields, separated by a comma. From Table II, we can then draw the following 4 observations:

1) the 'PApp' column does not correspond to port numbers in the sense of the transport layer (4) of the ISO model. In fact, some values are greater than 65,535. 2) LAMBERTII X & Y coordinates are identical for the same site name ('SiteName' column) 3) for the same site, we can have different antenna identifiers ('LocInf' column) because one site can serve several radio antennas 4) data were captured at regular 30-minute intervals 5) the 'df_U.csv' file contains all raw data from sites whose name begins with the letter 'U'.

A. Data Pre-Processing

This parsing phase, consisting of the *By-Site* and *By-Sector* aggregating steps, aims to reduce the amount of raw data to be processed with unsupervised MLAs, while retaining the underlying structure of the stream to enable better understanding and deeper analysis. Parsers and tools used in these steps are freely available on our repository [11]. These parsers have all been written in Perl which is a scripting language with advanced text processing and analysis functions, as well as a powerful regular expression engine. Perl also features string parsing capabilities, making it a particularly effective language for processing large data files.

1) *By-Site Aggregation*: In this stage, raw data is checked, cleaned, enriched and then aggregated by site. This analysis is based on the various fields described in Section III: (i) if the number of fields is different from 11 (either a field is missing or an additional field is detected), the line is deleted (ii) application type ('PortApp' field) is converted to an application group ('AppGroup' field) (iii) 'nPktUp' and 'nPktDn' fields are summed to obtain the values in the 'Packets' column (iv) raw data are aggregated by 'TimeSlot' and site (BTS) using the following 2-tuple key: 'TimeSlot; SiteName'. The numerical characteristics of the flows exchanged ('Duration', 'Users' and 'Flows') are summed, so that all information is retained. The 'PortApp' field is transposed into the 'AppGroup' field using the dictionary detailed below. This was generated from a correspondence table supplied by Orange, the mobile operator: 0) Unknown 1) Web 2) P2P 3) Download 4) CloudStorage 5) Mail 6) DB 7) Others 8) Control 9) Games 10) Streaming 11) Chat 12) VoIP 13) MailOperator 14) VPN 15) VVM 16) MMS 17) StreamAVSP 18) Portal.

At the end of this first aggregation stage, the raw data have been merged by site into flows, common application ports have been converted into a single application group and unused or

no longer required fields have been removed. Table III is a snippet of by-site aggregated data.

TABLE III
SNIPPET OF AGGREGATED DATA BY SITE

TimeSlot	SiteName	X/Y	AppGrp	Pkts	Dur.	Users	Flows
19-03-16 12:30	INVAL	939 / 477	0	712	193	3	4
19-03-16 12:30	INVAL	939 / 477	1	150095	20741	635	2193
19-03-16 12:30	INVAL	939 / 477	10	275357	11763	211	1747
19-03-16 12:30	INVAL	939 / 477	11	40595	30550	63	177
19-03-16 12:30	INVAL	939 / 477	12	6631	21	1	4

Data extracted from 'l.csv' file

The amount of data to be analyzed has been reduced from 184GB to 1.2GB, which represents a reduction rate close to 160. Total number of sites obtained is equal to 2,736.

2) *By-Sector Aggregation*: This second step merges the flows aggregated by site in the previous step into streams, to show network activity at sector level. This aggregation by sector means that we need to retain details of activity by application group, to enable more in-depth analysis in the next phase. In addition, we need to determine which sites are adjacent to each other, in order to group them into sectors.

This merging process was implemented using the Perl GEOHASH library [12]. Using the 'Coord_X' and 'Coord_Y' fields previously transposed into the GPS system, we were able to obtain the Geohash for each site. The length of the Geohash thus generated is called the *Precision*. We then merged the data from all the sites with identical Geohashes. Indeed, as explained in Section II-A2, sites with identical Geohash are neighbors. We therefore employed an aggregation key in the form of a 3-tuple key constructed as follows: 'TimeSlot; Geohash; App-Group'. The calculated data are transposed into the following form: 'TimeSlot,Geohash,Packets,Duration,Users,Flows'.

After this second stage of data aggregation, we obtain a single file describing user activity as a function of time ('TimeSlot' attribute) and by sector, the size of which depends on the 'Accuracy'. For the reasons given in the Section II-A2, we chose to create sectors by encoding the coordinates of each site in a 6-character Geohash and merging the data from sites with the same Geohash, which gave us 1,233 sectors of dimensions 1.2km by 0.6km as detailed by Table I. Table IV is an extract of the 'sectors_6.arff' file obtained after the sector aggregation step with precision equals 6.

TABLE IV
SNIPPET OF DATA AGGREGATED BY SECTOR

TimeSlot	Geohash	AppGroup	Packets	Duration	Users	Flows
2019-06-06 16:00	u09wj0	7	91506	1818461	183	183
2019-06-06 16:00	u09wj0	8	17133	10827	51	55
2019-06-06 16:30	u09wj0	18	8339	610	7	36
2019-06-06 16:30	u09wj0	3	4822262	103490	440	1083
2019-06-06 16:30	u09wj0	4	1624903	117040	518	1278

Data extracted from 'sectors_6.arff'

B. First Approach

Once generated, this 'sectors_6.arff' file is loaded into the Weka software for analysis. *Waikato Environment for Knowledge Analysis* toolbox [13] is an open-source software package developed by the University of Waikato in New Zealand that provides a comprehensive collection of visualization tools and

algorithms for data analysis and predictive modeling. All easily accessible via relatively intuitive graphical interfaces.

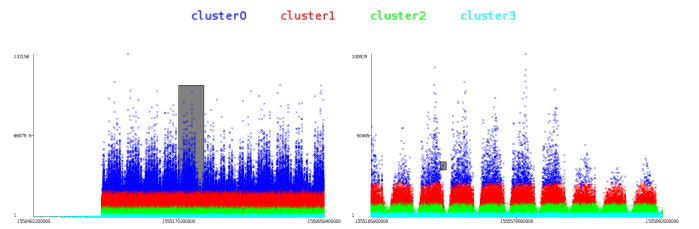
A first approach to analyze these streams or data aggregated by sectors was to employ the X-Means algorithm to understand how they are distributed and try to extract their underlying structure. X-Means allowed us to classify sectors on the basis of data characterizing user activity. To carry out this first clustering analysis, we chose to use the Euclidean distance as a metric, because it offers the best performance in terms of computation time for virtually identical cluster distribution results, as detailed in Table V. We can notice that the X-Means algorithm distributes data across 4 clusters whatever the metric employed.

TABLE V
DISTRIBUTION OF DATA ACROSS CLUSTERS

Distance	Cluster				Calculation time (s)
	0	1	2	3	
Euclidean	1%	7%	28%	64%	45
Chebyshev	1%	8%	29%	62%	47
Manhattan	1%	7%	27%	65%	49
Minkowski	1%	7%	28%	64%	66

Total number of points = 2,997,893

Then, still using WEKA, we project the 4 resulting clusters in 2 dimensions, with the 'TimeSlot' field on the x-axis and the 'Flows' field on the y-axis. These projections are represented by Figs. 3a and 3b. The Fig. 3a represents user activity for the entire *Île-de-France* over three months (12 weeks). We can see that this week's maximum points are higher than those of the previous week, as well as those of the following week. The Fig. 3b is a detailed view of the third week of April 2019 (Saturday 13/04/2019, Sunday 14/04/2019, ..., Monday 22/04/2019). We focused on April 2019 because some large-scale events, i.e. with huge crowds, took place during this period, and the first day of April 2019 was a Monday, which makes it easier to extract data by whole weeks.



(a) 12 week (3rd week of April in gray) (b) April 2019 (15/04/2019 in gray)

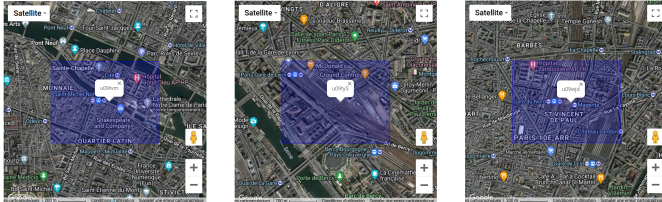
Fig. 3. Users activity for all sectors

Since we were looking to isolate large-scale events, we selected isolated extreme points in the *Cluster0* for Monday April 15, 2019. Data from some of these points are listed in Table VI. These data highlight three sectors with a higher level of activity than the others: *Notre-Dame de Paris* and two stations, *Gare de Lyon* and *Gare du Nord*. These areas are defined by their respective Geohashes, *u09tvm*, *u09ty5* and *u09wjd*. The Geohashes representing these sectors are projected (Fig. 4) and materialized by Figs. 4a to 4c.

TABLE VI
SNIPPET OF EXTRACTED OUTLIERS

TimeSlot	Geohash	AppGroup	Packets	Duration	Users	Flows
2019-04-15 16:00	u09ty5	8	17133	10827	51	55
2019-04-15 17:00	u09wjd	9	18979	7453	23	98
2019-04-15 16:00	u09wjd	7	3157	47859	76	326
2019-04-15 16:30	u09wjd	3	239790	7893	35	62
2019-04-15 22:00	u09ty5	1	91506	741461	221	141
2019-04-15 22:00	u09tvm	3	1712	1818461	183	183
2019-04-15 22:00	u09ty5	1	16285	1297180	2	4
2019-04-16 00:30	u09tvm	2	72511	10827	51	55

Data extracted from 'sectors_6.arff' file



(a) Notre-Dame de Paris (b) Gare de Lyon (c) Gare du Nord

Fig. 4. Some particular sectors with high activity level

The facts of : (i) be able to group data into different clusters (ii) to observe that a modulation of activity was visible when projecting into the data plane, reinforced our decision to employ an unsupervised MLA with the aim of determining this cluster distribution and compare it week by week.

To validate our "Unsupervised Anomalies Knowledge Flow" framework, we focused on the month of April, from Monday the 1st of April to Sunday the 28th of April, 2019. The analyzed streams relating to this period were split by week using the bash command 'grep' with regular expressions to filter time slots ('TimeSlot' field) and extract only the data relating to the sectors we wished to analyze ('Geohash' field).

C. Digital Signatures Extraction

This phase is carried out using the *X-Means* algorithm. This unsupervised MLA is used for aggregating detailed sector data and extract what we have termed *digital signatures* [14].

A digital signature is computed for each week and for each sector. Figure 5 represents allocation of network streams in each of the clusters, with the application group on x-axis and the number of users on y-axis, for different sectors or weeks.

From Figs. 5a and 5b, we can deduce that *Gare de Lyon* signatures are similar for these two weeks. This is also the case for the first and fourth weeks of April 2019. So, we can say that the signatures for the 4 weeks of April 2019 for the *Gare de Lyon* sector are identical. We can conclude from this that there were no anomalies and therefore no particular events occurred in this sector in April 2019.

The same conclusion can be drawn from Figs. 5c and 5d about *Gare du Nord*. The distribution and spread of data across the different clusters is identical for weeks 1 to 4, meaning that the digital signatures are identical.

A more interesting observation is that the signatures for *Gare de Lyon* and *Gare du Nord* are the same whatever the week studied. Group assignments are identical for the Figs. 5a

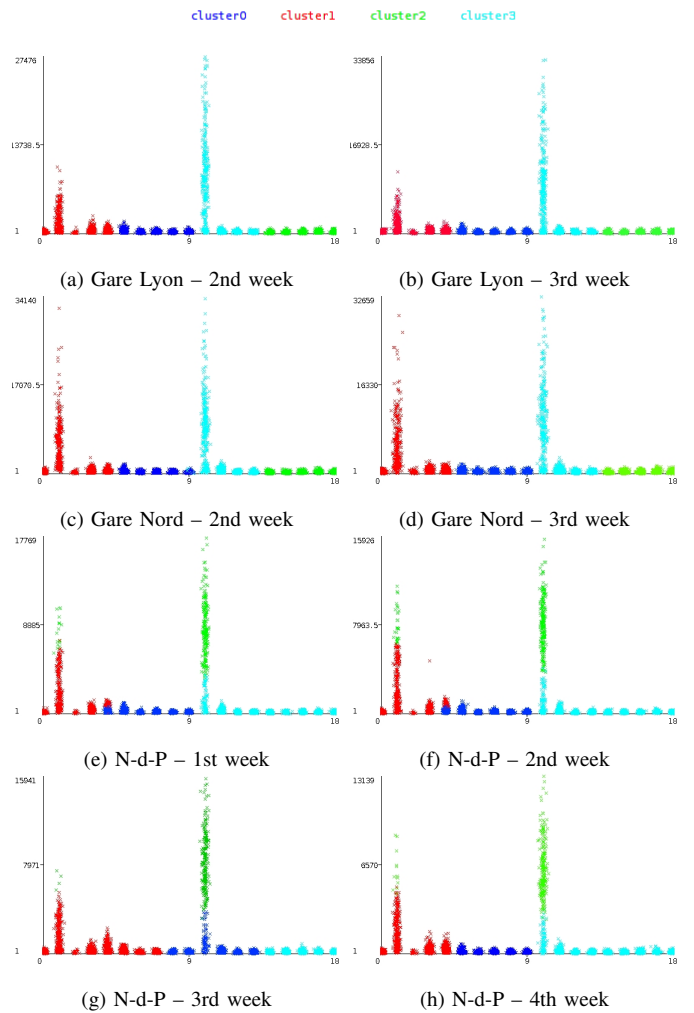


Fig. 5. Extracted DiSi with x-axis='AppGroup' & y-axis='Users'

to 5d. With our 'DiSi' concept, we seem to be able to classify sectors by type of activity (station, stadium).

The Figs. 5e to 5h describe the *Notre-Dame de Paris* sector. The signatures for weeks 1, 2 and 4 are uniform. An exception can be found in the Fig. 5h presenting the clusters of week 4, where an insignificant evolution appears with regard to the distribution of the 4 'AppGroup' clusters.

However, focusing on Fig. 5g, the signature is very different from the other digital signatures. AppGroups 5 (Mail), 6 (DB), 7 (Others), 10 (Streaming), 11 (Chat), 12 (VoIP) and 13 (MailOperator) are not assigned to the same clusters. This points to a change in network behavior that requires further analysis. This analysis is carried out in the last step using another unsupervised MLA.

D. Outliers detection

The final phase of our methodology is based on the LOCAL OUTLIER FACTOR algorithm. The WEKA implementation of this algorithm requires a class attribute to be able to represent them graphically. We therefore added an "Outlier" class and marked all instances as "False". It's also possible to configure

the distance measured by it during the grouping process. We selected *Chebyshev's* distance as it gave the best results, i.e. it allowed us to obtain the most outliers while still having a homogeneous distribution.

The LOF algorithm enabled us to extract around 15 outliers from the 5,075 instances contained in the third week of April, which represents 0.03%. On the other hand, only 2 outliers were extracted from the 5,098 instances in the first and fourth weeks of April, representing 0.06%. Some of these fifteen outliers linked to the third week are listed in Table VII.

TABLE VII
OUTLIERS EXTRACTED BY THE LOF ALGORITHM

Instance	TimeSlot	AppGroup	Packets	Geohash	Flows
528	2019-04-15 18:00:00	12	1745	u09tvm	9
714	2019-04-15 23:30:00	12	360	u09tvm	2
808	2019-04-16 02:30:00	4	8347	u09tvm	68
926	2019-04-16 07:00:00	4	9425	u09tvm	175
379	2019-04-27 18:00:00	15	2354	u09wmd	6
443	2019-04-27 20:00:00	10	8256608	u09wmd	8550
559	2019-04-27 23:00:00	5	43482	u09wmd	135
74	2019-05-12 22:00:00	16	167652	u09wmd	605

E. Anomaly correlation

From Table VII, we can deduce that an anomaly occurred from April 15, 2019 at 6pm to April 17, 2019 at 10am in the *Notre-Dame de Paris* area. This anomaly can be correlated to a real event and is related to *Notre-Dame de Paris* fire. Another anomalies occurred in April 27th and May 12th that correspond to French soccer cup final match and the 'Metallica' concert in *Stade de France*.

V. CONCLUSION & FURTHER WORKS

From this study, we can say that our framework named "Unsupervised Anomalies Knowledge Flow" enables the detection of network anomalies. In our use case, the anomalies are linked to a change in user behavior, echoing the different types of mobile application used at different times in particular geographical areas. We chose to detect this type of behavior because mobile application activity was one of the features available in our dataset. By extension, our approach can be transposed to other networks presenting flows with different characteristics, another capture time periods or a sector aggregation method based on other notions such as subnet mask or link type, for example. We can say that our approach is *generic*, *cyclic* and *fractal* because it is adaptable, iterative and more or less detailed depending on the size of the constituted sectors.

Our framework consists firstly in *aggregating data into sectors* representing specific areas to extract *network activity features*. Based on these features, we compute a *digital signature* for each particular period (weeks) to highlight networks behavior. Then, for each sector, we compared the signatures to detect any changes in network behavior to highlight anomalies. The final step enabled us to detect outliers, establish correlations and confirm detected anomalies.

Our "Unsupervised Anomalies Knowledge Flow" is able to detect other anomalies and events such as the French Football Cup final match between Rennes and Paris on April 28, 2019,

or the 'Metallica' concert on May 12, 2019, held at the *Stade de France*. In addition, we can classify sectors according to the type of activity. This idea needs to be confirmed by selecting more sectors to compare. We plan to apply our concept to smaller sectors defined by 8-character Geohash values, in order to detect less important events or types of activity. Another study would focus on the ability to detect a particular lack of network traffic when a high level of activity is the norm or required. We also plan to study how activity is distributed across the mobile network by analyzing smaller Geohash values.

Further work could involve a hybrid approach combining supervised [15] and unsupervised machine learning algorithms to detect network security anomalies or implement a real-time knowledge flow.

ACKNOWLEDGMENT

This work was partially supported by the ANR CANCAN (ANR-18-CE25-0011) and CoCo5G (ANR-22-CE25-0001) projects [10], [16]. We thank MIKE TINNEY, ALEXANDRE KARIM from ESIEE, CHI-DUNG PHUNG from Cnam, CEZARY ZIEMLIKI, and ZBIGNIEW SMOREDA from Orange Labs, for their data collection support.

REFERENCES

- [1] M. S. Parwez, D. B. Rawat, and M. Garuba, "Big Data Analytics for User-Activity Analysis and User-Anomaly Detection in Mobile Wireless Network," vol. 13, no. 4, pp. 2058–2065.
- [2] E. S. C. Vilaça, T. P. B. Vieira, p. u. family=Sousa, given=Rafael T., and p. u. family=Costa, given=João Paulo C. L., "Botnet traffic detection using RPCA and Mahalanobis Distance." IEEE, pp. 1–6. [Online]. Available: <https://ieeexplore.ieee.org/document/8896228/>
- [3] S. B. Wankhede, "Anomaly Detection using Machine Learning Techniques," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pp. 1–3.
- [4] M. Szmit, R. Wezyk, and M. Skowroński, "Traffic anomaly detection with Snort."
- [5] O. Podzins and A. Romanovs, "Why SIEM is Irreplaceable in a Secure IT Environment?" pp. 1–5.
- [6] C. Maudoux and S. Boumerdassi, "Network Anomalies Detection by Unsupervised Activity Deviations Extraction," in *2022 Global Information Infrastructure and Networking Symposium (GIIS)*, pp. 1–5.
- [7] Geohash Intro — Big Fast Blog. [Online]. Available: <https://web.archive.org/web/20120112004608>
- [8] D. Pelleg and A. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters."
- [9] Detect Outlier (LOF) - RapidMiner Documentation. [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/cleansing/outliers/detect_outlier_lof.html
- [10] Content and Context based Adaptation in Mobile Networks. Agence nationale de la recherche. [Online]. Available: <https://anr.fr/Project-ANR-18-CE25-0011>
- [11] Cmaudoux / digital-signatures-extraction. [Online]. Available: <https://bitbucket.org/cmaudoux/digital-signatures-extraction/src/master/>
- [12] Geohash - Great all in one Geohash library - metacpan.org. [Online]. Available: <https://metacpan.org/pod/Geohash>
- [13] Weka 3 - Data Mining with Open Source Machine Learning Software in Java. [Online]. Available: <https://www.cs.waikato.ac.nz/ml/weka/>
- [14] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools*, ser. MK Series in Data Management Systems. Morgan Kaufmann.
- [15] C. Maudoux, S. Boumerdassi, A. Barcello, and E. Renault, "Combined Forest: A New Supervised Approach for a Machine-Learning-based Botnets Detection," in *2021 IEEE Global Communications Conference (GLOBECOM)*, pp. 01–06.
- [16] CoCo5G Project – Collecte de trafic, analyse contextuelle et conception de fonctions basées sur des données 5G / 5G network data analytics. [Online]. Available: <https://coco5g.roc.cnam.fr/>