ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Unsupervised Real-time Anomaly Detection for Multivariate Mobile Phone Traffic Series

Evelyne Akopyan[1], Angelo Furno[1], Nour-Eddin El Faouzi[1], Eric Gaume[2]

[1] Univ. Gustave Eiffel, Univ. Lyon, ENTPE, LICIT,
F-69518, Lyon, France

[2] Univ. Gustave Eiffel, GERS, LEE,
F-44344, Bouguenais, France

Abstract. Real-time anomaly detection in urban areas from massive data is a recent research field with challenging requirements. This paper presents a lightweight framework for real-time anomaly detection in multivariate time-series extracted from large-scale Mobile-phone Network Data (MND). Our solution relies on unsupervised machine learning applied to MND collected at individual antennas of a nation-wide French mobile phone network operator. The proposed framework is based on a two-step approach: (i) the offline stage aims at assessing the typical behaviour of the antennas; (ii) the online stage performs real-time comparison of incoming data with respect to the detected typical behaviour. Results related to a real case-study of terrorist attack in the city of Lyon show that our framework can successfully detect an emergency event almost instantaneously and locate the anomalous area with high precision.

1 Introduction

Large-scale user mobility data has recently started to be used to detect anomalous mass behaviour at district or city-wide scales, and solutions leveraging such data to promptly inform relevant authorities of critical situations have been proposed [1]. Global Positioning System (GPS) data are among the most used type of mobility data leveraged to such purpose because of their high spatial accuracy and temporal resolution. However, this type of data is rarely available on large scales and users can be easily identified from the tracks, raising relevant privacy concerns. On the other hand, Mobile-phone Network Data (MND), passively collected by network providers for billing or network management purposes, is an opportunistic form of user mobility data which has the advantage of being easily anonymized and massively provided on large scales and extremely high penetration rates [2]. However, such data are often regarded as less accurate than GPS data, thus difficult to analyse without proper pre-processing steps.

Anomaly detection is a well-studied topic in many fields including fraud detection, security breaches, natural disaster alert, etc. A thorough review of existing anomaly detection techniques for temporal data is provided in [3, 4]. More recently, a particular interest has raised towards performing real-time anomaly detection in urban areas [5] because of the growing availability of massive online data from Internet of Things (IoT) devices, social networks data as well as highly instrumented urban infrastructures. Urban areas need indeed efficient real-time monitoring systems to ensure the security of large masses of human

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

population. The anomaly detection system introduced in [6] is among the few examples leveraging MND to identify the time and location of anomalous mass behaviour and the geographical spread of the anomalies. However, the authors' solution relies on supervised machine learning, thus requiring anomalous data to be explicitly labelled, which is hard to achieve on large spatio-temporal scales. Such label constraint is not present in the solution proposed in [7] that exploits an unsupervised approach, thus reducing the effort and costs necessary to acquire labels for the training process. However, the solution still relies on a computationally-intensive deep learning framework which requires powerful hardware support and makes it difficult to frequently update the trained model. The approach is therefore less adapted to a fully distributed city/region-wide massive deployment of the anomaly detection system.

The main contribution of this paper is a real-time unsupervised lightweight anomaly detection framework for large-scale MND. It requires very few parameters to be set, and is based on a statistical model that ensures lightweight computation and almost instantaneous anomaly detection, which makes our solution a perfect fit for an online deployment and particularly adapted to monitor densely populated areas. Sec. 2 describes the overall architecture of the framework. The first results for a real-situation case study are presented in Sec. 3.

2 Anomaly Detection Framework for MND

2.1 MND Dataset Description

The proposed approach is specifically designed to detect anomalous situations from MND. For each antenna of an instrumented mobile network, the provider is able to probe and collect traffic information related to different types of *services*. Traffic *volumes* are thus provided, per-service and per-antenna, with a one-minute temporal granularity. Our solution aims at detecting, in real-time, anomalies in traffic volumes with respect to a compressed description, which we call *signature*, of the typical traffic volume observed per-service at each antenna of the monitored region. In this paper, for prototyping and evaluation purposes, we only use a historical MND dataset containing 3G and 4G mobile data collected from March 2019 to June 2019 on the nation-wide Orange network. 50% of the data is used for training, and the rest is used for the real-time anomaly detection stage. Note that the online stage of the proposed solution is designed to work on streaming data that will be provided, in real-time, by the network provider in the final operational deployment. Each entry of the available dataset is defined as a vector $s_a(t) = \langle v_a^0(t), ..., v_a^{n-1}(t) \rangle$ where a is the identification code of the antenna, t is the time of the observation and v_a^i is the volume of events observed for service i. Tab. 1 details the fields of the dataset.

2.2 Framework Architecture

Fig. 1 depicts the architecture of the proposed framework. First of all, the framework includes a pre-processing module (M.0) performing the following operations on the MND: (i) multiple antennas can be located the same geographical coordinates. In such case, they are treated as one single *node* and related volumes

Field	Value
t	Year (2019)-Month (3-6)-Day (1-31) Hour:Minute
a	Latitude, Longitude

i	3G: {Call, SMS, Packet-Switched, Circuit-Switched}
	4G: {Call, SMS, Handover, Service Request}

v_a^i	Number of events in $\mathbb{N}$

Table 1: MND set fields and values

are aggregated; (ii) volumes can be unavailable in case of intermittent demand. In this case, a zero volume is assumed at the corresponding time slot.

Then, the architecture is composed of the following main modules: (M.1) an offline machine learning stage that extracts a *typical* per-service traffic behaviour at each node, as detailed in Sec. 2.3; (M.2) a real-time anomaly detection stage based on the comparison of the real-time data with the typical behaviour of the node, as detailed in Sec. 2.4.

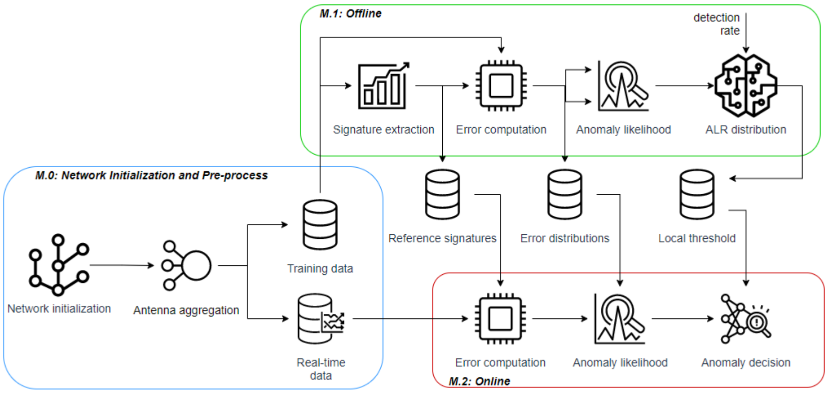


Fig. 1: Framework architecture with the M.0, M.1 and M.2 modules

2.3 Offline module (M.1): Training Stage

The methodology for building reference signatures for mobile data is based on our previous work [8] and is the first step to extracting a typical behaviour. The signatures are defined per-service and per-node. For each pair of node-service, all data is grouped by the chosen level of granularity. As studied in our previous work, a weekly signature with a granularity of 1 minute (finest granularity for the available dataset) appears adequate because it allows grasping the most relevant differences in traffic consumption, by relying on the well-known hypothesis of cyclic nature of human activities [9]. To build the signature for a given node a and service i, we consider the service volume data $v_a^i(d,t)$ per minute for a specific day d over multiple days of past observations, *i.e.*, $d \in \mathbf{d} = \{d_1, d_2, ...d_M\}$. Let us denote as $\mathbf{d}^\delta \subset \mathbf{d}$ the set of days in the dataset that correspond to the day of the week δ. For instance, $\mathbf{d}^{\text{MON}}$ groups all Mondays in the dataset. Then, the generic element in the signature of node a for service

i is defined as:

$$r_a^i(\delta, t) = \mu\left(\left\{v_a^i(d,t) \mid d \in \mathbf{d}^\delta\right\}\right)$$

for time slot t during day of the week δ, where $\mu(\cdot)$ represents the median operator, applied to the set within parentheses. Unlike the mean, the median is robust to outlier values, which occur frequently in mobile traffic data possibly due to sudden variations in human communication activities and anomalous events. The latter must be filtered out in order to derive a reference normal behaviour. The weekly signature $r_a^i(t)$ is then defined as the concatenation of the one-minute time-ordered samples $r_a^i(\delta, t)$ over the seven days of the week.

The weekly signature $r_a^i(t)$ is however still sensitive to inherent noise in the data. Therefore, to further refine it, a low-pass filter is applied to the time series for smoothing purposes. The spectral density of the weekly signature is analysed to choose the right cut-off frequency for the filter: once converted to the frequency domain with Fast Fourier Transform (FFT), only the highest power frequencies are kept. After applying inverse FFT, the resulting filtered signal $\tilde{r}_a^i(t)$ corresponds to the reference signature of the given node a for service i.

MND are an intermittent kind of data, *i.e.,* the observed traffic can be null for some services at given moments due to the absence of network activity, thus raising the problem of infinite values for many error metrics such as the relative error. Therefore, we use the instantaneous absolute error (AE) with respect to the reference signature, defined at each time slot t as: $\epsilon_a^i(t) = |v_a^i(t) - \tilde{r}_a^i(t)|$, where v_a^i is the real-time volume for service i at node a.

In the offline phase, the AE is leveraged to fit a model of the error at each node and for each service. To that purpose, we compute the AE with respect to the reference signature over a large set of days of past traffic observations, used as a training dataset. Let us define X_a^i the random variable associated to the AE of service i at node a. By using the training dataset, the distribution of X_a^i is thus fitted to a Gamma law Γ_a^i using the Maximum Likelihood Estimator method. A comparative analysis performed on the available dataset with multiple theoretical distribution laws has confirmed the Gamma one being the most adequate choice to accurately fit the error distribution and detect rare events, such as abnormally high errors. We assume that the variables $X_a^{i \in I}$ of node a are independent: this hypothesis has been verified by a correlation test between the values of AE computed for each service individually.

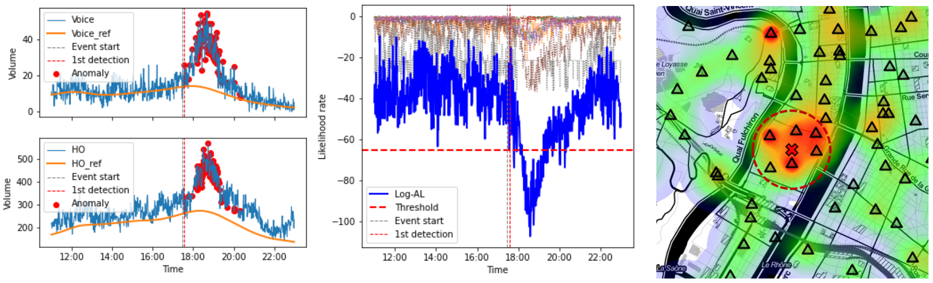
Based on the fitted error distribution, we define the anomaly likelihood rate $\mathcal{L}$ at time t for service i at node a as the tail distribution at point $\epsilon_a^i(t)$, *i.e.,* $\mathcal{L}_a^i(t) = P(X_a^i \geq \epsilon_a^i(t))$ where the probability P is derived from the fitted Gamma law. An anomaly is observed at time t on a node a if the observed traffic related to the ensemble of network services significantly deviates from the reference behaviour. To that purpose, we need to convert the multivariate anomaly likelihood indicator into an univariate one. Therefore, we define a global Anomaly Log-likelihood Rate (ALR) at time t for all services $i \in I$ as the sum of the logs of the likelihood rates, *i.e.,* $\Lambda_a(t) = \sum_{i \in I} \log \mathcal{L}_a^i(t)$. The system detects an anomaly at time t when the ALR is lower than a threshold value, *i.e.,* $\Lambda_a(t) \leq \theta_a$. The θ_a threshold value is defined as the q^{th} quantile of the ALR distribution fitted for that node over the training dataset. q is the only hyperparameter of our framework and can be selected by fixing an average detection rate. It is worth to highlight that θ_a is node-specific and is stored for the real-time detection step.

2.4 Online module (M.2): Real-Time Detection

After the training phase, at each node a, a vector $s_a(t)$ of real-time data is fed to the model. By comparing the volumes of this vector with the reference signatures $\tilde{r}_a^i(t)$, the model derives an AE value $\epsilon_a^i(t)$ for each service i. The AE values are then used to determine whether the real-time vector $s_a(t)$ is consistent with the typical behaviour or can be declared as an anomaly. The model derives the ALR value at time t as described in Sec. 2.3 by using the typical Gamma distribution of each service. Then, if the ALR value is lower than the threshold θ_a, the real-time vector $s_a(t)$ is labelled as an anomaly. Otherwise, $s_a(t)$ is labelled as normal. The time complexity for one detection is $\mathcal{O}(nm)$ where $n = |I|$ and m is the time complexity of the distribution fitting operation. Note that the absolute value of the ALR can be used as an indicator of the intensity of the anomaly.

3 Experimental Results: Terrorist Attack in Lyon, France

A terrorist bombing occurred at Bellecour square, Lyon (France), on Friday 24th May 2019 at 5:28 PM[1]. In the following analysis, we focus on nodes in a radius of 300 m around the attack location. During the training stage, we set $q = 1/720$ for each node, *i.e.,* a detection rate of one anomaly every twelve hours.



(a) Real-time volumes (b) ALR evolution (c) Anomaly area

Fig. 2: Event Detection on node-scale and city-scale

Fig. 2a depicts the event detection from the real-time volumes of the call and handover services at one of the nodes in the selected area. The first anomaly is detected at 5:34 PM, *i.e.*, six minutes after the official time of the attack. Fig. 2b shows, for the same node, the evolution in time of the likelihood rate of each service, as well as the ALR (blue curve). We can see that soon after the bombing occurs, the ALR rapidly falls below the threshold. Concerning the whole selected area, our method detects a first anomaly for multiple nodes

[1] *Le Progrès*, 2020 [online] https://www.leprogres.fr/actualite/2020/05/27/un-an-apres-l-attentat-a-lyon-les-riverains-de-la-rue-victor-hugo-ont-ils-tourne-la-page

between 5:34 PM and 6:13 PM. The larger offset with respect to the official time of the attack for some nodes could be explained by considering antennas' network coverage features. Some of them could cover zones that are still geographically close, but not in the immediate proximity, of the location where witnesses were mostly concentrated at the beginning of the event. These results are encouraging because other nodes, farther away from the attack location, do not detect any anomaly at all on that day, thus making the anomaly easy to locate with accuracy at the scale of the city as shown in Fig. 2c (anomalous zones are colored in red).

4 Conclusions and research directions

In this paper, we introduced a new MND-based framework for anomaly detection in real-time. We have proposed an efficient algorithm based on simple and unsupervised machine learning concepts: our model relies on very few parameters and is lightweight enough to be deployed on large scales such as cities. The model has also proven to be accurate and very reactive, hence minimizing the delay of detection in a real-situation case study. We aim to more thoroughly study the false alert rate and delay of detection as well as the impact of adding contextual data from social networks on the reliability of the system.

Acknowledgements

This work is supported by the French ANR research projects DISCRET (grant number ANR-19-FLJO-0002-01) and PROMENADE (grant number ANR-18-CE22-0008).

References

[1] James P Bagrow, Dashun Wang, and Albert-Laszlo Barabasi. Collective response of human populations to large-scale emergencies. *PloS one*, 6(3):e17680, 2011.

[2] Kashif Sultan, Hazrat Ali, and Zhongshan Zhang. Call detail records driven anomaly detection and traffic prediction in mobile cellular networks. *IEEE Access*, 6, 2018.

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[4] Manish Gupta, Jing Gao, Charu Aggarwal, and Jiawei Han. Outlier detection for temporal data. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 5(1):1–129, 2014.

[5] Huichu Zhang, Yu Zheng, and Yong Yu. Detecting urban anomalies using multiple spatio-temporal data sources. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–18, 2018.

[6] Adrian Dobra, Nathalie E Williams, and Nathan Eagle. Spatiotemporal detection of unusual human population behavior using mobile phone data. *PloS one*, 10(3), 2015.

[7] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[8] Angelo Furno, Marco Fiore, Razvan Stanica, Cezary Ziemlicki, and Zbigniew Smoreda. A tale of ten cities: Characterizing signatures of mobile traffic in urban areas. *IEEE Transactions on Mobile Computing*, 16(10):2682–2696, 2016.

[9] Andres Sevtsuk and Carlo Ratti. Does urban mobility have a daily routine? learning from the aggregate data of mobile networks. *Journal of Urban Technology*, 17(1):41–60, 2010.